Author:
Abdelrahman, Hend Abu Elmakarem

Title:
Phylogenomic and genomic analyses of the metamonad Pseudotrichomonas keilini, a free-living anaerobic relative of Trichomonas

Phylogenomic and genomic analyses of the metamonad *Pseudotrichomonas keilini*, a free-living anaerobic relative of *Trichomonas*

By

Hend Abu-Elmakarem

University of BRISTOL

School of Biological Sciences
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Masters by Research (MScR) in the Faculty of Science.

October 2020

Word count: fifteen thousand,eight hundred twenty eight

Abstract

The origin and evolution of eukaryotes has been linked to the rise of oxygen following the Great Oxidation Event, but anaerobic habitats are common today, have existed throughout the Earth's history, and are rich with eukaryotic life. The Parabasalia are an ancient anaerobic lineage, and the most speciose lineage of Metamonada, a major lineage of eukaryotes. The most well-studied Metamonads are parasites including *Trichomonas vaginalis* and *Tritrichomonas foetus*,and *Giardia intestinalis* but very little genome data is available for free-living members of the group. Here, we sequenced the genome and transcriptome of *Pseudotrichomonas keilini*; a free-living metamonad.

Comparative genomic analysis indicates that *P. keilini* possesses a metabolism and gene complement that are in many respects similar to its parasitic relative *T. vaginalis*. These similarities include a hydrogenosome (anaerobic mitochondrial homologue) that we predict to function much as in *T. vaginalis*. They also include a complete glycolytic pathway that is likely to represent one of the primary means by which *P. keilini* obtains ATP. Phylogenomic analysis indicates that *P. keilini* branches within a clade of parasitic parabasalids, consistent with the hypothesis that different parabasalid lineages evolved towards parasitic or free-living lifestyles from an endobiotic, anaerobic or microaerophilic common ancestor.

Acknowledgements

Working on this project has been a great learning curve for me. It involved a lot of ups and downs with moments of self-doubt. I would not have been able to go past it and challenge myself if it were not for the ultimate guidance and support that I received from my supervisors.

I would like to thank Dr. Tom Williams for being the supportive, and incredible supervisor that he is, for entrusting me with taking the lead in this project, and encouraging me to try everything without the fear of failure. He has given me the opportunity to be part of a great team of researchers which taught me a lot through the different meetings and workshops. I am also grateful for the effort he exerted which allowed me to receive the funding to start this degree.

To Dr. Celine Petitjean, I will be forever indebted to the remarkable academic and emotional support you have been giving me. You never hesitated to give me the time to explain and tackle the different research problems I faced. You have also helped me grow as a researcher and have given me the confidence to go beyond where I dared to dream of going. I owe you what I have learned in research planning, scientific communication and presentation skills. Thank you!

I also wish to express my deepest gratitude to Dr. Christopher Kay for his help in planning and conducting several experiments in this project. You have been of great help!

Starting this degree would not have been possible without the funding I received from the School of Biological Sciences to waive the tuition fees. I would like to thank everyone that facilitated this process.

My special thanks go to my aunt and mentor Dr. Maha Abdelnasser. I would not have made it this far without your constant support, guidance, and unconditional love.

I am incredibly lucky for having the supportive family and friends which were always present when I needed them despite the distance. Thank you for believing in me, for giving me a hand when I needed one, and for not sparing time, effort, or resources to make my time in Bristol easier.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Hend Abu-ELmakarem DATE: 31/07/2020

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.0.1 Oxygen and the origin of eukaryotes

Oxygen only reached its current level in the atmosphere 450 million years ago[117, 177, 109] due to the origination of land plants [177, 109] while eukaryotes are estimated to have risen around 1.6 Ga ago [98, 152]. We do not know very much about how free-living eukaryotes thrive without oxygen, today or in the distant past. While we cannot go back in time to see how that happened, we can rather study present anaerobic eukaryotes which can shed lights on how their ancestors grew, divided and replicated in anoxic environments.

It is still debated how early eukaryotes managed to live, survive and divide with the low oxygen levels present at their time. There are two main views about that; the first, is that the early eukaryotes were obligate aerobes but were able to adapt to hypoxic and anoxic conditions through lateral gene transfer (LGT) [9, 14, 96, 78, 173, 170, 171, 116, 115]. Supporters of this view argue that eukaryotes were unable to survive hypoxia throughout their evolutionary history and hence they gained access to anaerobic environments by gaining genes through lateral transfer. Views about the last eukaryotic common ancestor (LECA) are that it was unable to survive anaerobiosis and rather gained that ability late after the divergence of the main eukaryotic lineages by LGT from prokaryotic anaerobes [14, 96, 78, 173, 170, 171, 116, 115, 16].

The second view is that the first eukaryotes were facultative anaerobes and that the mitochondria they harbored were also facultative anaerobes [136, 181, 125, 126, 123, 132, 196, 127]. In that case, eukaryotes were able to survive anaerobic environments through vertical evolution. Moreover, anaerobic energy metabolism enzymes are not restricted to anaerobic eukaryotes as they can be found in aerobic ones as well such as algae [12] and *Naegleria gruberi* whose genome revealed the presence of enzymes that were thought to be restricted to anaerobic eukaryotes only [67].

1.0.2 Anaerobic eukaryotes

Many protists or unicellular eukaryotes occupy environments where free O_2 is scarce or not available at all like in anaerobic sediments [152, 65, 19, 176, 15]. Moreover, various ciliate species have an ensosymbiotic relationship with methanogenic archaea that live within them [57, 54].

These anaerobic eukaryotes use different enzymes for energy production. Such as pyruvate :ferridoxin oxidoreductase (PFOR or PFO) and FeFe-hydrogenase (hydA) which were first indentified in the hydrogenosomes of trichomonads [121].

> Hydrogenosomes are organelles related to the mitochondria which can be found in anaerobic eukaryotes. These organelles lack a complete Kreb's cycle and do not use the Electron Transport Chain (ETC) to produce energy. Instead, they produce ATP along with molecular Hydrogen by the action of several enzymes such as FeFe-hydrogenase and pyruvate:ferridoxin oxidoreductase (PFOR or PFO) [136].

PFO which is an analogue to the mitochondrial enzyme pyruvate dehydrogenase (PDH), oxidizes pyruvate to produce Acetyl-CoA, reduced ferredoxin (Fd^-) and CO_2. H_2 gas is then produced by the reoxidation of Fd^- by hydA which converts electrons from pyruvate to protons [139, 175]. Molecular phylogenies have shown that some of the mitochondrial genes were detected in these organelles, supporting the hypothesis that both mitochondria and its related organelles were a result of the same endosymbiotic event that gave rise to the canonical mitochondria.

There have been two contrasting views on the age of anaerobic eukaryotes in regards to the mitochondrial-related organelles that they contain. Organisms containing mitochondria-like organelles such as diplomonads, and parabasalids were thought to be older in the eukaryotic tree and forming early branches, it was later revealed that that was a case of 'long-branch attraction' in which molecular sequences that have undergone huge numbers of evolutionary changes tend to cluster together in phylogenetic trees.

> That clustering of fast-evolving sequencing can be falsely interpreted as them being 'deeper' in phylogenetic trees, or older in the tree of eukaryotes; however, the eukaryotes lacking the canonical mitochondria are in fact fast-evolving species and not older ones.

1.0.3 The origin of Mitochondria

The full story of how the mitochondria emerged and was integrated into the eukaryotic cell as an organelle is yet to be revealed. However, it is generally accepted that

mitochondria emerged by a merging event of an endosymbiotic alphaproteobacterium and a host cell closely related to Asgard Archaea. To become a fully integrated organelle in the eukaryotic cell, several evolutionary events and changes have occurred including gene losses, gene gains by horizontal gene transfers (HGT), metabolism and reproduction integration, and membrane transporters insertion. It is still unclear what exactly that original endosymbiont alphaproteobacterium was like and when exactly it was integrated into the eukaryotic cell. [157]

Commonly known as the powerhouse of the cell, mitochondria are double-membraned organelles found in most eukaryotic cells with a function to produce energy in the form of adenosine triphosphate (ATP). The main mechanism by which energy is produced is aerobic respiration which is basically the oxidation of pyruvate to CO_2 from which reduced forms of co-factors are produced which fuel the electron transport chain (ETC) through which ATP is produced with oxygen as the last electron acceptor.

There are many biochemical features of mitochondria aside from the ATP production. These features include the metabolism of nucleotide, amino acid, and quinone. In addition to the synthesis of proteins, the catabolism of fatty-acids, steroids and lipids. Mitochondria also play an important part in the bio-genesis of the iron-sulfur cluster (Fe/S), in the ion homeostasis, and cell apoptosis. All of the above are examples for the mitochondrial functions besides aerobic respiration.

[37, 73, 11, 145, 114]

The hypothesis that mitochondria originated from an alphaproteobacterial symbiont was proposed by Mereschowsky in 1905 [133]. Several scientists have greatly supported it such as Margulis in 1967 [159]. As controversial as it was to propose that at that time, phylogenetic analyses since the 1970s and 1980s have been supporting that hypothesis as they show that these organelles were different from the eukaryotic lineage and rather belong to a prokaryotic one [29, 165, 192]. Recent phylogenetic analyses also give great support to that hypothesis [186]; however, it is still unclear which lineage is the closest relative within that group[155]. Regarding the provenance of the host lineage that was involved in the endosymbiosis event, recent phylogenomic analyses indicate that it was most closely related to a newly discovered Archaeal group called the Asgards [169, 194]. Investigations of the cell biology and genomic analyses of different protists and multicellular eukaryotes show that their last eukaryotic common ancestor (LECA) was a mitochondrion-containing ancestor referred to as the mitochondrial cenancestor. [156, 58, 45]. The mitochondria in LECA continued to evolve in different directions. The discovery and analysis of the lineage of anaerobic eukaryotes reveal that there are several forms of mitochondria or

mitochondrial-related organelles that evolved from LECA, which helped them adapt to living in low oxygen environments[74]. The form of mitochondria they possess have significant differences compared to the widely known aeorbic form. Another misconception about the mitochondria is that it is present in all eukaryotic cells. Nevertheless, there is only one known eukaryote that completely lacks any form of mitochondria or any related organelle which is *Monocercomonides* [103]. All of that has given us more clues about the kind of mitochondria that was present in LECA. Along with having mitochondria as a fully integrated organelle, LECA's mitochondria was capable of aerobic respiration in addition to other biochemical functions that were discovered in modern mitochondria [74, 185, 178]. Furthermore, other analyses have revealed similarities between mitochondrial sequences with different groups of alphaproteobacteria [1, 180].

It is still unclear which type of alphaproteobacteria the pre-mitochondrial endosymbiont originated; however, phylogenomic analyses link it to the group of Rickettsiales which contains both endosymbiotic and parasitic bacteria [186, 185] which has also been supported by gene conactinations from complete genomes [160, 10, 188]. Moreover, analyses relying on genes encoded on both the mitochondrial genome and the nuclear genome, recover mitochondria within the Rickettsiales, being sister to Anasplamataceae, Rickettsiceae and Midichloriaceae[186, 185]. In other cases, it was suggested that mitochondria form an independent deep branch in the tree of alphaproteobacteria [155]

There are several proposals that picture the first endosymbiotic mitochondrial ancestor as biochemically versatile capable of being a facultative aerobic photosynthetic bacterium. This would allow its host to move and live in aerobic niches[190], both oxidize sulfide which is produced by host respiration [166] and release hydrogen as product of fermentation [123], or by producing organic photosynthate [64, 40].

Given the complexity of the biochemical capabilities of the mitochondrial cenancestor, it has been argued that the endosymbiont and host metabolic association were multifaceted [71]. Another hypothesis was that the relationship between the pre-mitochondrial alphaproteobacterium and the host cell was non-mutualistic as the bacterial cell aggressively invaded the host [47, 48].

another hypothesis was discussed in the Imachi et al.(2020) paper [97] in which they isolated an arhaeon

> at the interface of eukaryotes and prokaryotes which supports a prokaryotic origin for eukaryotes.

In their paper, they proposed a new eukaryogenesis model called the entangle–engulf–endogenize model, also referred to as the E^3 model. It describes the relationship between the aerobic endosymbiotic organotrophic pre-mitochondrion and the host archaeon.

Their newly discovered archaeon suggested alternative route in which the archaeal host engulfed the metabolic partner with the use of extracellular structures while forming a primitive chromosome-surrounding structure that is similar in topology to the nuclear membrane; however, this conjecture needs further evidence to support it. The Imachi model for eukaryogenesis can be described in the following six steps:

i) The host archaeon which is of syntrophic/fermentative nature is hypothesised to degrade amino acids to short-chain fatty acids and H_2, it is possible that it was able to do so by interacting with H_2 and O_2-scavenging Sulfate-reducing bacteria SRB. ii) The archaeon host may have then interacted with a partner of a facultatively aerobic organotrophic nature which is the pre-mitochondrial cell, which was able to scavenge toxic O_2, the interaction with SRB may have continued as it was not essential but could have been beneficial. iii) The interaction of external structures between the archaeon host and the aerobic partner (pre-mitochondrion) by mechanical or biological fusion in order to enhance physical interaction and engulf the partner for concurrent development of a primitive nucleoid-bounding membrane and endosymbiosis. iv) The continuation of the archaeon host and pre-mitochondrion symbiont interaction after engulfment. v) Development of ADP/ATP carrier (AAC) by the pre-mitochondrion endosymbiont while the initial direction of ATP transport still not clear. vi) Endogenization of pre-mitochondrion partner symbiosis by the archaeon host by delegation of ATP generation and catabolism to the pre-mitochondrion endosymbiont and establishment of an ATP channel from symbiont to host.

1.0.3.1 Becoming an Organelle

For the pre-mitochondrial alphaprotteobacterium to become a fully integrated organelle in the eukaryotic cell, several changes must have taken place for it to have the biochemical properities enabling it to produce ATP along with biosynthetic and metabolic pathways. Such changes include; 1) the insertion of transporter/carriers molecules into the inner membrane of the symbiont, 2) origination of the protein-import machinery, 3) genome reduction and loss of redundant genes or unnecessary ones, 4) endosymbiotic gene transfer (EGT) to become integrated in the nucleus, 5) cell envelope modification, 6) biochemical pathways and systems integration between host and symbiont, 7) origination of a mechanism for organelle division, 8) cristae speciation, 9) contact sites evolution between the endomembrane and proto-mitochondria system, 10) re-targeting and localization of proteins of diverse origins to mitochondria, 11) anchors evolution between cytoskeleton and mitochondria [157].

Membrane transporters are one of the key elements of the mitochondria. Such transporters like the mitochondrial carrier family (MCFs) could have evolved from a

single carrier that was integrated into the inner membrane of the proto-mitochondria [40, 10, 8].

1.0.3.2 Mitochondrial Proteomes

Mitochondrial proteomes are chimeric [74, 178, 71, 101] with a size of approximately 1,000 proteins [37, 73, 11, 145, 114]. Recent evidence traces only 10-20% of the mitochondrial proteome to an alphaproteobacterial origin [185, 72]. And about 20-30% show proteobacterial affinity [178, 72]. Even more interesting, around 40% of the mitochondrial proteome cannot be traced to either prokaryotic or viral origins [178]. It is likely that these proteins with no prokaryotic origin evolved within eukaryote groups after LECA [74, 178]. However, the reamianing 15% of the proteome has prokaryotic affinity with non-proteobacterial homologs[178]. It is likely that these 15% were laterally transferred before the endosymbiosis to the pre-mitochondrial endosymbiont, or came from archaeal origin in the proto-eukaryotic host, or genes that came from later gene transfers from bacteria or viruses to the nucleus of the proto-eukaryote. During the process of organellogenesis, the proto-mitochondrial compartment lost its alphaproteobacterial identity through events of gene losses and gene gains in the proto-eukaryotic genome [71, 101, 72, 112].

Many of the proteins with alphaproteobacterial origin whether in the mitochondrial or nuclear genomes, have roles in aerobic respiration [185, 72]. These aerobic respiration functions include: 1) the ETC to make ATP through chemiosmosis, 2) the translation of genes encoded in the mitochondrial genome through the mitochondrial ribosome, 3) Krebs cycle that provides reduced forms of (NADH and $FADH_2$) to the ETC, 4) production of acetyl-CoA by the oxidative carboxylation of pyruvate which then starts the Krebs cycle, 5) the β-oxidation pathway of fatty acids which provides NADH for Krebs cycle or the respiratory chain, 6) biosynthesis of several co-factors such as the Fe/S clusters, biotin and heme which are necessary for the assembly of several proteins of the respiratory complexes in addition to other mitochondrial enzymes, 7) the biosynthesis of ubiquinone and cardiolipin which are required for a proper functioning respiratory chain [185, 72].On the other hand, many of the mitochondrial proteins which come from a eukaryotic origin have important functions in the inner and outer membranes of mitochondria such as protein import, organelle division, metabolite transport and others [178, 72].

In the initial phase of becoming an integrated organelle in the eukaryotic cell, proto-mitochondrial genomes had been through several changes which resulted in having a greatly reduced genome compared to the endosymbiont genome. This reduction phase probably started when the mitochondrial symbiont was unable to replicate outside the host cell [129, 128].

1.0.3.3 Mitochondrial-related organells (MROs): Hydrogenosomes

Until recently, it was thought that some eukaryotes completely lack any form of mitochondria, especially for eukaryotes such as microsporidians like *Trachipleistophora* and *Vairimorpha*, others like *Giardia*, *Trichomonas* and *Entamoeba*. The hypothesis that some ancestors of eukaryotes referred to archezoa, never contained mitochondria was originated by Cavalier-Smith. As the genomes of some archezoa contain orthologs for mitochondrial proteins such as Hsp60 or Hsp70, it was suggested that these organisms descended form ones once containing mitochondria. [56] It was later discovered that many of these organisms contain a mitochondrial-related organelle (MRO) of some kind. The only exception up to date is *Monocercomonides* which completely lacks any form of energy-production organelles [103, 102]. Following the discovery of mitochondrial organelles with different metabolic capabilities, Müller et al.(2012) [139] have suggested a classification of mitochondria and its related organelles that evolved from mitochondria into five forms; 1) the classic aerobic mitochondria in which Oxygen is the terminal electron acceptor in oxidative phosphorylation such as that of *Homo sapiens*. 2) Anaerobic mitochondria; which is found in anerobic organisms for which Oxygen is not used in ATP production like in *Ascais summ*. 3) Hydrogen-producing mitochondria which contains sub-units for proton-pumping complex I, and synthesizes ATP via anaerobic pyruvate metabolism which is found in *Blastocystis sp.*. 4) Hydrogenosomes which lack a complete Krebs cycle and electron transport chain and produces energy anaerobically with H_2 and CO_2 in the process which were thoroughly described in *Trichomonas vaginalis*. 5) Mitosomes which completely lack genomes and do not produce ATP which is the case in *Giardia intestinalis*.

The Figure 1.1 illustrates the wide diversity of MROs across the tree of eukaryotes. It is notable how the Excavata supergroup is rich in the different evolved forms of mitochondria such as hydrogenosomes and mitosomes. Hydrogenosomes are double-membaraned organelles that evolved from mitochondria. While the main function of the mitochondria is producing ATP by electron transport chain and re-oxidizing NADH produced in glycolysis, hydrogenosomes lack electron transport chain system, but they produce ATP and hydrogen gas in the process.

They were first discovered in the early 1970s at Rockefeller University by D. G. Lindmark and M. Müller where hydrogenosomes were observed in *Tritrichomonas foetus*, which is a cattle parasite [121].

Hydrogenosomes are found in anaerobic eukaryotes including unicellular ones like ciliates and trichomonads, some fungi also harbour it and even animals [46]

In an attempt to understand the origin of hydrogenosomes, several studies have been focusing on the evolution of certain key enzymes that were found in the hy-

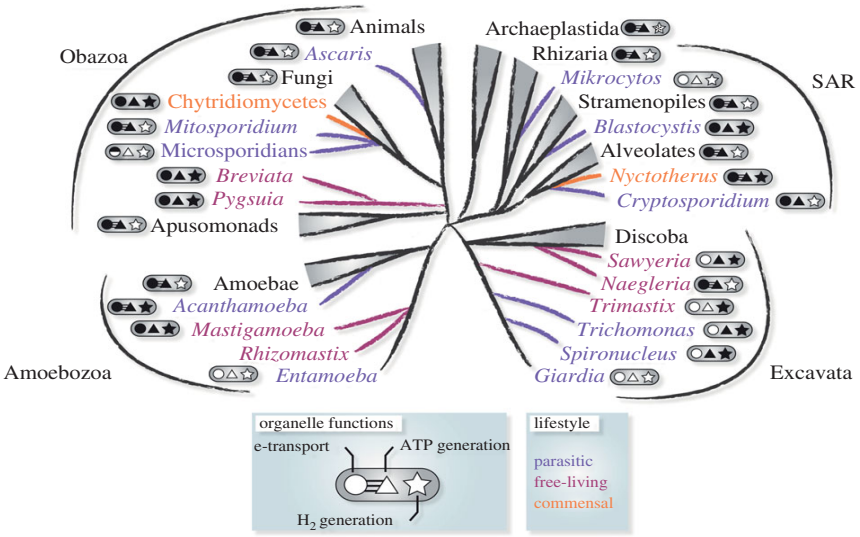

Figure 1.1: Distribution of MROs across the major supergroups of eukaryotes. Organisms with parasitic (purple), commensal (orange) or free-living (red) lifestyles are indicated. Metabolic functions of each organism's MRO are indicated: shaded shapes represent the presence of electron transporting complexes (circle), ATP synthesis (triangle) and hydrogen production (star). Where at least one proton-pumping complex (CI, CIII, CIV) and ATP synthase (CV) was identified, the circle and triangle are joined by three lines. '*' represents algal lineages where hydrogen production is located in the plastid. Figure modified from Stairs et al. (2015) [172]

drogenosomes of all organelles. These enzymes include [FeFe]-hydrogenase(hydA), and its three maturase enzymes; small GTP-binding protein (hydF), FeFe-hydrogenase assembly protein (hydG) and radical SAM domain containing protein (hydE). In addition to pyruvate:ferredoxin oxidoreductase (PFO), Succinyl coenzyme A synthetase (SCS), and the Iron–sulfur cluster (ISC) pathway enzymes such as Iron-sulfur cluster assembly enzyme (IscU), Cysteine desulfurase (IscS), Fxns,[2Fe–2S] Fdxs, Isa2, Nfus, Grx5, and Ind1s as well as chaperone Hsc20.

It was thought that these enzymes were restricted to the eukaryotes that do not have a conventional mitochondrion; however, it was later discovered that they can be found in a wide variety of eukaryotes including humans [59].

Since hydrogenosomes can be found in several unrelated organelles across the tree of eukaryotes, that suggests that these organelles have evolved independently several times. As it is now accepted that hydrogeneosmes first evolved from mitochondria, the evidence for that being the similarity of the protein import machinery in both organelles. This is not entirely different from the mitochondria since most of the mitochondrial genes are encoded in the nuclear genome of the host. [59]

The similarities between the mitochondrial and hydrogenosomal protein import machinery can be seen through the extensive analysis of the *Trichomonas vaginalis* hydrogenosome. It can be seen that N-terminal target sequence is carried by the hydrogenosomal ferridoxin into the hydrogenosome of *Trichomonas* which is the same targeting sequence found in the yeast mitochondria [151].

Since they evolved from mitochondria; MROs share a lot of the biochemical pathways and subunits which are found in the mitochondria. So far, the iron-sulfur cluster (ISC) pathway is the only conserved biochemical pathway in all forms of mitochondria and MROs which could indicate a significant importance for the survival of the cell [59, 157]. Furthermore, hydrogenosomes have similar protein import machinery and contain many ancestrally mitochondrial proteins such as mtHsp70 and mtHsp60. These proteins from both mitochondria and hydrogenosomes group together in phylogenetic trees [182]. Hydrogenosomes tend to lack a genome, perhaps because of the lost Oxidative phosphorylation and rather uses the proteins that are synthesized in the cytosol with *Nyctotherus ovalis* being the only organism whose organelles retained parts of its genome. The debate hinges on the origins of the genes involved in anaerobic metabolism that are specifically found in hydrogenosomes.

1.0.3.4 The origin and evolution of PFO and hydrogenase enzymes

In the hydrogen hypothesis, it was proposed that two prokaryotic cells interacted together to start the eukaryogenesis. It was proposed that the host was archaebacterium in need of hydrogen, for which it integrated in it an alphaproteobacterium endosymbiont which was the ancestor of the mitochondria which provided the host

with hydrogen. The result of this integration was the evolution of the eukaryotic cell which has genes from archaebacterial origin and others from eubacterial one, mainly the ones responsible for energy metabolism such as PFO and [Fe-Fe]-hydrogenase [124].

As previously mentioned, key hydrogenosomal enzymes such as PFO and [Fe-Fe]hydrogenase were found in eukaryotes that do not contain hydrogenosomes. For example, they were detected in *Spironuclues*, *Giardia* and *Entamoeba* [89, 143]. In addition, genes encoding for hydrogenase in different organisms were cloned, such as *Trichomonas vaginalis*[34, 89], *Neocallimastix frontalis*,*Nyctotherus ovalis* [5] and *Piromyces sp.* [49, 184]. What was common among all of them was that all of these genes encode the iron-only [Fe-Fe] hydrogenases which is not found in archaebacteria, and is rather found in eubacteria.

There is considerable variation in the structure of these enzymes. For instance, *Piromyces sp.*, *Neocallimastix frontalis*, and *Trichomonas vaginalis* have a longer form of [Fe-Fe] hydrogenase as it contains four accessory iron-sulfur [FeS] clusters at the N-terminus which is the case in some eubacterial enzymes[90]. In addition to that, *Trichomonas* contains two 'short-form' [Fe-Fe] hydrogenases with absent first [4Fe4S]-cluster and terminal [2Fe2S]-cluster.

While it was previously hypthesized that the PFO and [Fe-Fe]-hydrogenase enzymes were acquired form the mitochondrial endosymbiont [60, 123], phylogenetic analyses do not support that hypothesis. Instead, phylogenetic trees for PFO in several studies show the sequences from eukaryotes grouping together in one clade. These results are consistent with the hypothesis that PFO was present in early eukaryoutes [158, 91]. Moreover, phologenetic trees of the [Fe-Fe]-hydrogenase enzyme show that the eukaryotic and alphaproteobacterial sequences of the enzyme do not cluster together which concludes a non-endosymbiotic origin of the enzyme [59]. All of the above can indicate the origin of the hydrogenosomal enzyme PFO and [Fe-Fe]-hydrogenase to be as old as the origin of the eukaryotic cell itself.

1.0.4 Phylogenetic relationships in Excavata

To find a consensus on the phylogeny of excavates and their interrelationships, several studies have been conducted. One of these studies is the phylogenomic analysis by Moreira and colleagues (2007) where they constructed phylogenetic tree of eukaryotes and used a combination of subunit ribosomal DNA (SSU rDNA) and large-subunit (LSU) rDNA sequences. Their analysis showed that Excavata, are polyphyletic in the analysis of the SSU rDNA tree but are monophyletic in that of the LSU tree. They also showed that their result was the same even after including the fast-evolving groups [135].

Another study which was done by Hampl et al. (2009), suggested the monophyly of Excavata after the removal of genes and species that are rapidly evolving Fig 1.2. However, recent revised studies have shown otherwise. Adl et al. (2019) have published a revised classification of eukaryotes where they proposed that Excavata is comprised of three sub-clades; Metamonada, Discoba, and Malawimonada. They further showed the uncertainty about the relationships between these sub-clades and among the other eukaryotic clades referred to as *Incertae sedis Eukarya*. They have further stated that monophyly is lost in Excavata while it was present in metamonads and discoba separately [77, 2].

The controversial interrelationships between the sub-clades of Excavata can be shown through a recent study by Heiss and colleagues (2018). Their phylogenomic analysis which included the recently discovered malawimonad *Gefionella okellyi* n. gen. n. sp. showed that metamonads form a clade with malawimonads and not discoba. [82] which can be shown in Figure 1.3 .

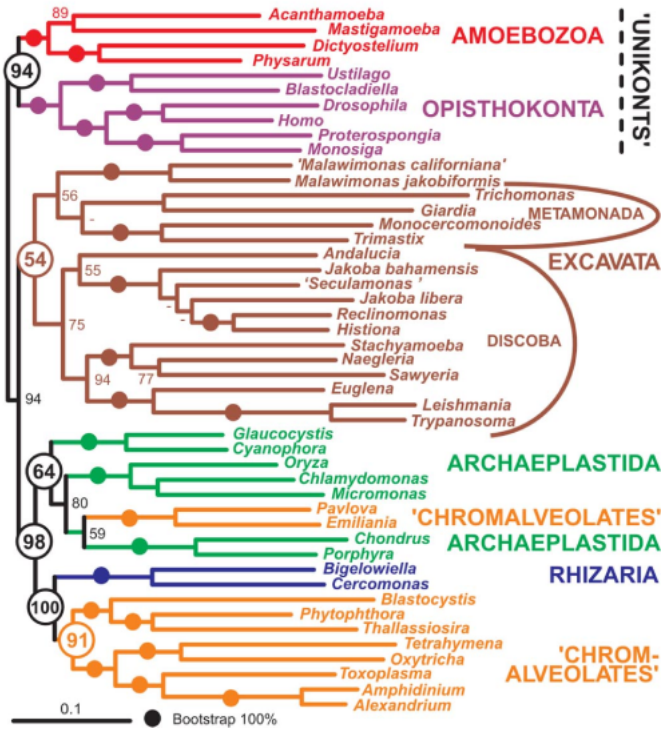

Figure 1.2: A maximum likelihood tree after the removal of 1,750 of the longest-branch gene sequences. The tree was constructed in RAxML using the WAG $+\Gamma$ model. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping. Branches that received maximum support by all methods are indicated by full circles, dashes indicate bootstrap values $<50\%$. Although the analyses did not assume a root, the tree is displayed with the basal split between "unikonts" and bikonts as suggested in ref. Figure modified from Stechmann et al. (2003) [174] [77]

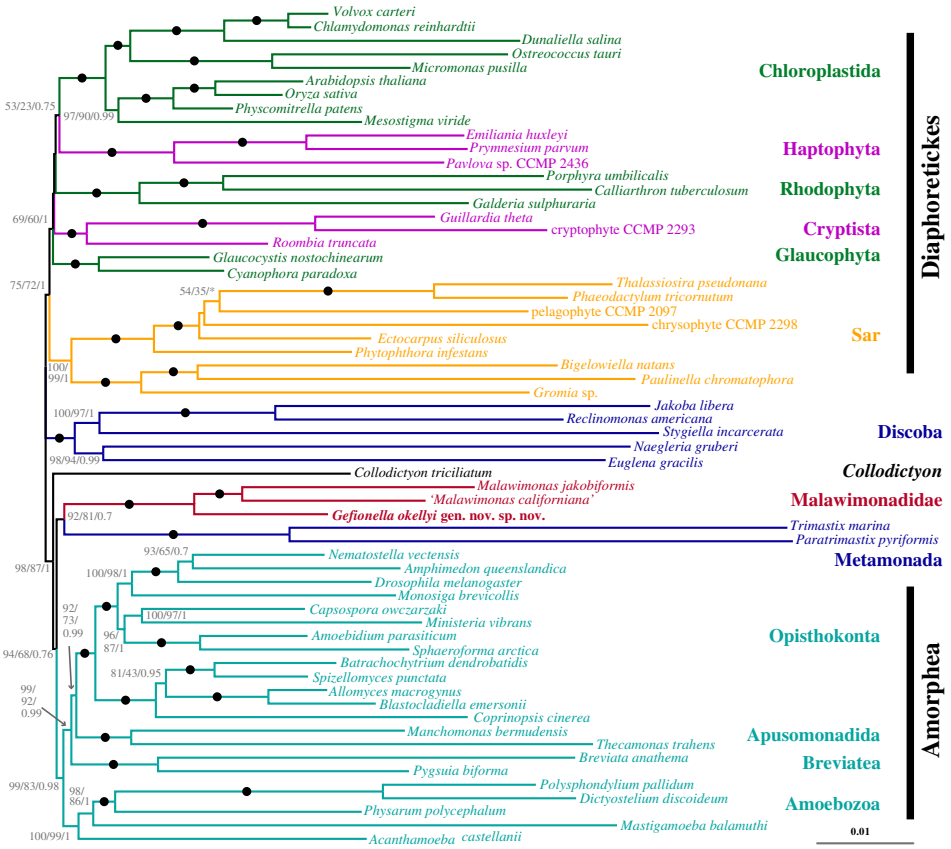

Figure 1.3: Phylogenetic tree of eukaryotes, based on 159 genes, with 23,000 fast-evolving sites and 22 long-branch taxa excluded from initial dataset. Tree shown was inferred under LG+C60+Γ4+F model of sequence evolution using IQ-TREE. Statistical support values are, in order: LG+C60+Γ4+F model UFboot from IQ-TREE, LG+Γ4+F model BP from RAxML, and CAT-GTR+Γ4 model Bayesian posterior probabilities from PhyloBayes-MPI. Filled circles represent maximal support (i.e. 100/100/1). Unlabelled branches received less than 50% UFboot support. Asterisks denote branches that were not recovered in inferred phylogeny for a given analysis. Figure adapted from Heiss et al. (2018) [82].

1.0.5 Parabasalids and their position in the eukaryotic family

Parabasalia represent a group of anaerobic single-celled eukaryotes which belong to the Metamonada phylum. They are mainly distinguished by their lack of canonical mitochondria. Instead, they harbor different mitochondrial-related organelles (MROs) such as the hydrogenosome or mitosome. The taxonomy of the group members has been debated over the years.

> Parabasalia were originally divided into two assemblages based on their morphology into trichomonads and the hypermastigotes.

The main difference between them is in the cell complexity and the number of flagella per mastigont. The mastigont system is composed of four main parts: the parabasal and sigmoid filaments from which the parabasalia group takes its name, the pelta–axostyle system which is made of microtubules, the costa, which is a periodic rootlet and other filaments [17].

While trichomonad cells are mostly simpler and generally have up to six flagella per mastigont, the hypermastigotes can have several thousands of flagella with complex cells. [31, 43, 87]. Ultrastructural similarities between the trichomonads and the hypermastigotes revealed by the electron microscope gave stronger support to the to unite the two groups into the parabasalia superorder [86, 84, 85, 179]. Figures 1.5,1.6 show these structural similarities as both of them contain axostyle, pelta and parabasal body; however, they are more transformed in hypermastigotes. All of that led to the proposal of the superorder Parabasalia by Honigberg(1973) [88].

> Despite the monophyly of Parabasalia as a superorder, molecular phylogenetic trees have shown that both assemblages trichomonads and hypermastigotes are not monophylotic on their own.

[197].

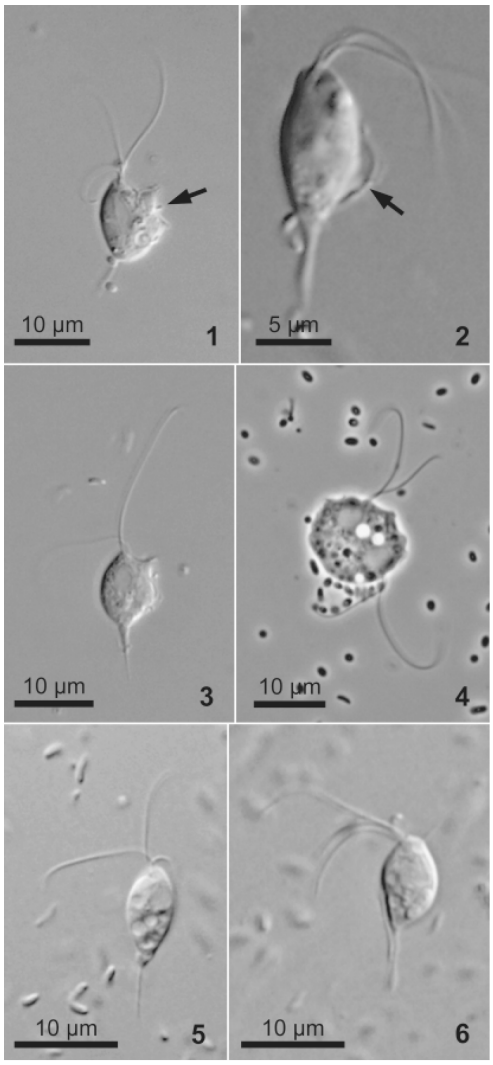

Figure 1.4: 1–6. Light micrographs of *Pseudotrichomonas keilini* NY0170 (Japan), *P. keilini* LIVADIAN (Cyprus), and *Lacusteria cypriaca n. g., n. sp.* (Cyprus). **1.** A cell of *P. keilini* NY0170 (Japan) with three anterior flagella and a posterior flagellum (arrow) along with an undulating membrane. **2.** A cell of *P. keilini* LIVADIAN (Cyprus) with three anterior flagella and a posterior flagellum (arrow) along with an undulating membrane. **3.** An immature cell of *P. keilini* NY0170 (Japan) with two anterior flagella. **4.** A dividing cell of *P. keilini* NY0170 (Japan) with two nuclei and two pairs of anterior flagella. This cell is slightly compressed and enlarged by a cover slip. **5.** A cell of *L. cypriaca* n. g., n. sp. (Cyprus) showing three anterior flagella and an axostyle. **6.** A cell of *L. cypriaca* (Cyprus) showing three anterior flagella and an undulating membrane. Figure adapted from Yubuki et al. (2010) [193]

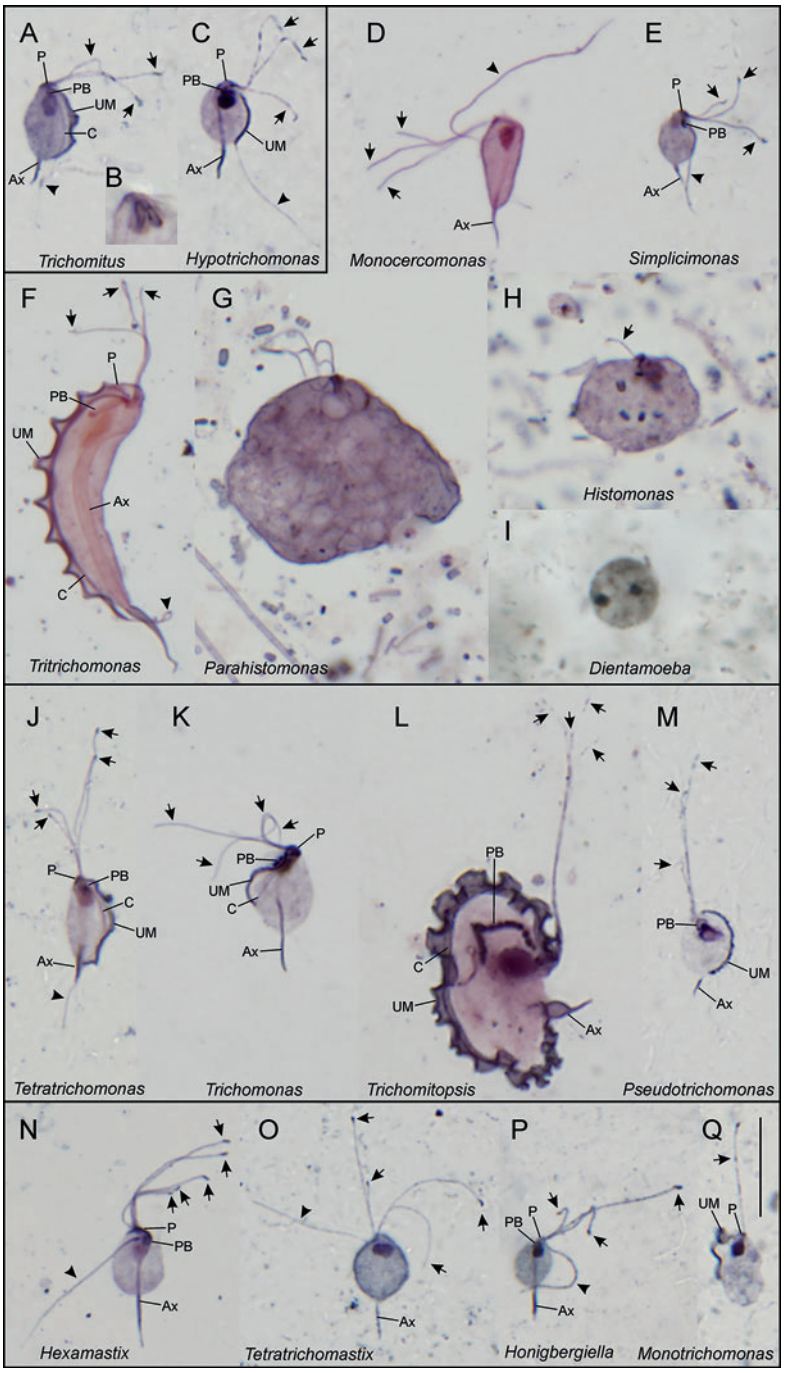

Figure 1.5: (Continued)

Figure 1.5 ***(previous page)*****: Light-microscopic morphology of trichomonads; Hypotrichomonadida (a–c), Tritrichomonadida (d–i), Trichomonadida (j–m), and Honigbergiellida (n–q)**. Protargol-stained cells, bright field. **(a)** *Trichomitus batrachorum* from *Bufo bufo*. **(b)** Parabasal body of *Trichomitus batrachorum* from *Testudo radiata*. **(c)** *Hypotrichomonas acosta* from *Leptopelis* sp. **(d)** *Monocercomonas colubrorum* from *Tropidophis melanurus*. **(e)** *Simplicimonas similis* from *Melamphaus faber*. **(f)** *Tritrichomonas augusta* from *Lacerta vivipara*. **(g)** *Parahistomonas wenrichi* from *Meleagris gallopavo*. **(h)** *Histomonas meleagridis* from *Meleagris gallopavo*. **(i)** *Dientamoeba fragilis* from *Homo sapiens*. **(j)** *Tetratrichomonas* sp. from *Macaca silenus*. **(k)** *Trichomonas tenax* from *Homo sapiens*. **(l)** *Trichomitopsis termopsidis* from *Zootermopsis angusticollis*. **(m)** Free-living *Pseudotrichomonas keilini*. **(n)** *Hexamastix coercens* from *Acomys* sp. **(o)** *Tetratrichomastix* sp., origin uncertain. **(p)** *Honigbergiella ruminantium* from *Bos taurus*. **(q)** Free-living *Monotrichomonas* sp. **Scale bar in Q** = 10 μm; it applies for the whole plate. **Labels**: *arrows* anterior flagella, *arrowhead* recurrent flagellum, *Ax* axostyle, *C* costa, *P* pelta, *PB* parabasal body, *UM* undulating membrane. Figure modified from Cepicka, Dolan, and Gile (2016) [197]

Although they are traditionally classified within the excavate supergroup [39], parabasalids lack the feeding groove and cytosome which are key features of the group. Instead, they possess a microtubular and non-microtubular cytoskeletal elements systems [167].

A revision of taxonomic classification of parabasalids was done by Cepicka et al.(2010) which divided parabasalia into six classes: Tritrichomonadea, Trichomonadea, Hypotrichomonadea, Cristamonadidea, Spirotrichonymphea, and Trichonymphea with eight orders and 17 families [42] which was adopted in a revised classification of eukaryotes [4].

Recent classification attempts for the group such as that by Cavalier-Smith(2013) divided parabasalia as a super-class composed of two classes, Trichomonadea and Trichonymphea. The former was then subdivided into subclasses Eotrichomonadea and Cristamonadea. Eotrichomonadea was further divided into two orders Tritrichomonadina and Trichomonadida. The latter has two suborders Trichomonadina and Honigbergiellina. While Cristamonadea was divided into two orders of Cristamonadida and Spirotrichonymphida. On the other hand, Trichonymphea was divided into two orders, Trichonymphhida and Lophomonadida. Current molecular evolution and phylogenetic studies show the paraphyly and polyphyly of several taxa of that system [41].

There are over 450 identified parabasalid species, the best-studied of which are mainly parasitic species that infect livestock and humans [3, 32, 42, 191]. They mainly moved from living in a lower intestinal tract environment to the genitourinary, respiratory or digestive tracks. An example of these species is *Trichomonas vaginalis*, *Trichomonas gallinae*, *Histomonas meleagridis* and *Tritrichomonas foetus*. [197]. The most medically important discovered parabasalid to date is *Trichomonas vaginalis*. It is known to be the causing agent for the most common nonviral STD; Trichomoniasis. It infects the male, and female genitourinary tracts of 187 million people every year around the world [131] with raising occurrences of HIV transmission[149].

The general shared characteristics in the group include a parabasal apparatus which is a Golgi complex connected with striated fibers, crypropleuromitosis which is a closed mitosis associated with an external spindle, in addition to hydrogenosomes.

1.0.5.1 Free-living parabasalids

A few free-living parabasalian species have been described such as *Monotrichomonas carabina*, *Lacusteria cypriaca*, *Ditrichomonas honigbergii*, *Honigbergiella sp.* and *Pseudotrichomonas keilini* [24, 26, 63, 18, 193]. Such organisms can be found in both freshwater and marine environment where there is no (anoxic) or very little Oxygen (microoxic) levels.

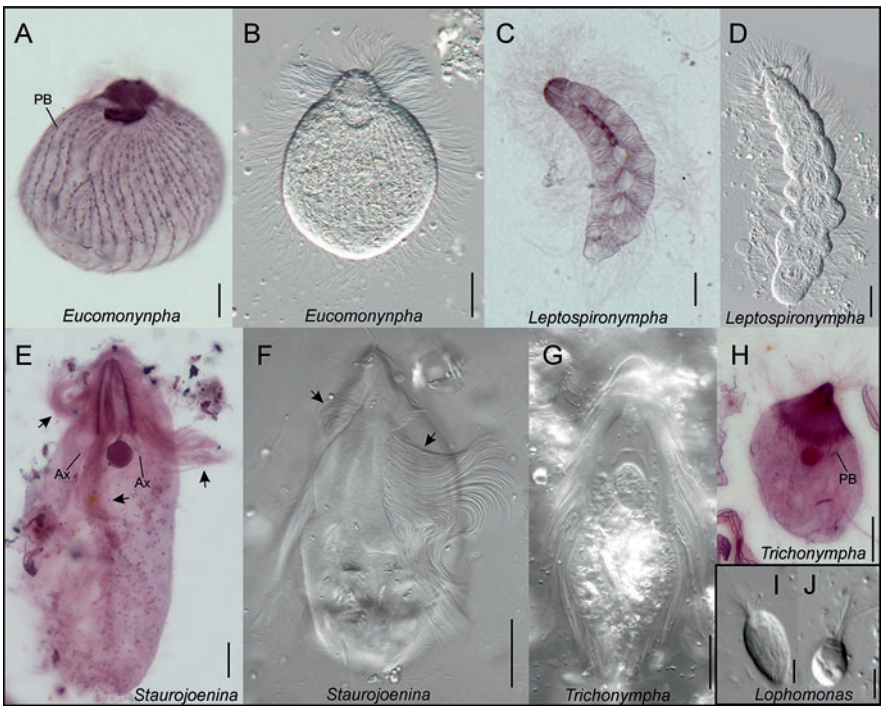

Figure 1.6: Light-microscopic morphology of the hypermastgotes; Trichonymphida (a–h) and Lophomonadida (i, j). **(a)** Protargol-stained *Eucomonympha* sp. from *Cryptocercus primarius*. **(b)** Living *Eucomonympha* sp. from *Cryptocercus primarius* observed under DIC. **(c)** Protargol-stained *Leptospironympha* sp. from *Cryptocercus primarius*. **(d)** Living *Leptospironympha* sp. from *Cryptocercus primarius* observed under DIC. **(e)** Protargol-stained *Staurojoenina* sp. from *Neotermes cubanus*. **(f)** Living *Staurojoenina mulleri* from *Neotermes jouteli* observed under DIC. **(g)** Living *Trichonympha* sp. from *Cryptocercus punctulatus* observed under DIC. **(h)** Protargol-stained *Trichonympha* sp. from *Reticulitermes flaviceps*. **(i)** Living *Lophomonas striata* from *Periplaneta americana* observed under DIC. **(j)** Living *Lophomonas blattarum* from *Periplaneta americana* observed under DIC. **Scale bars** = 10 μm for **a, c, i,** and **j**; 20 μm for **b, d, e, g,** and **h**; and 50 μm for **F**. **Labels**: *arrows* flagellar bundles of *Staurojoenina*, *Ax* axostyle/axostylar filaments, *PB* parabasal body. Figure modified from Cepicka, Dolan, and Gile (2016) [197]

However, little is known about the free-living ones . We are yet to understand their metabolism and energy production mechanism in anoxic environments. Phylogenetic relationships between the free-living parabasalids show that they do not form a clade. They instead group with their parasitic relatives in several different positions [193]. This work focuses on one free-living parabasalid, *Pseudotrichomonas keilini*.

Being the first free-living parabasalid discovered [24, 26], there have been several attempt to culture *Pseudotrichomonas keilini*, the first of which was deposited in the American Type Culture Collection with an ID (ATCC 50321). A small ribosomal subunit (SSU) rDNA was extracted from that culture and was available in GenBank (AY319274). Nevertheless, further morphological and phylogenetic analyses of SSU rDNA by Hampl and colleagues(2007) indicated that the culture in fact belongs to the genus *Honigbergiella* and resembles the species *Honigbergiella ruminantium* as it was lacking an undulating membrane which is a one of the main characteristics of *P. keilini* [79].

Dufernez and colleague (2007) reported that *Pseudotrichomonas* was found endobiotic in cattle [50], but this was later shown to be a related member of *Honigbergiella* [42]. Hence, the first actual culture deposited for *P. keilini* was deposited by Yubuki and colleagues (2010) [193]. In their study, the group isolated, cultured and sequenced the small ribosomal subunit (SSU) rDNA from two *P. keilini* strains. The samples were isolated from different locations, a marine mangrove sediment close to Ishigaki Island, Japan, and from Cyprus, close to Voroklini [193].

The two strains are named as follows: the first being from Japan was called *Pseudotrichomonas keilini* NY0160(Japan) which was deposited at the American Type Culture Collection (ATCC), Mansas, VA, USA with an accession number (PRA-328), and the second was called *P. keilini* LIVADIAN (Cyprus) which was then deposited at Charles University in Prague, Department of Parasitology, Prague, Czech Republic. The SSU rRNA genes of of both strains were sequenced three times to ensure the cultures were not contaminated with other parabasalids.

To better understand the evolutionary relationships among parabasalids and their other eukaryotic relatives, a clear classification of their position in the eukaryotic tree is needed. The Excavata super of Eukaryotes includes the Metamonada clade where Parabasalia belongs.

The relationship between parabasalids and their other eukaryotic relatives can be understood by looking at the Parabasalia position in the tree of Eukaryotes.
The relative subclades to Parabasalia are Prexaxostyla which comprises Oxymonada and trimastigids and Fornicata which includes Diplomonadida, Retortamonadida, and *Carpediemonas*-like organisms [167, 4, 195]. Phylogenetic analysis suggest that Prexaxostyla forms the deepest branch in the clade while Fornicata is a sister group

of parabasalia [80, 77, 105].

Since many parabasalids live in the gut of termites, the age of Parabasalia can be estimated in relation to the origin of termites which is about 150 million years ago in the Jurassic/Cretaceous boundary [134, 30]. Parabasalids existed before termites and although molecular clock analysis is not yet available for the group, the estimated age for Excavata ranges between 900 and 1.8 billion years [147, 61].

1.0.6 *Pseudotrichomonas keilini*

1.0.6.1 Taxonomic and Phylogenetic position

Subsequent to its discovery in 1935, Prof. Anne Bishop decided to place *P. keilini* (then called *Trichomonas keilini*) in the *Trichomonas* genus due to its high similarity to the other *Trichomonas* species described. Another flagellate was later discovered in France with similar characteristics and was thought to be *P. keilini*; however, it was placed in the genus "Eutrichomastix" instead [113]. Later, a revised taxonomic position was done by Bishop 1939 in which she showed that the morphology of the free-living flagellate does not fit in the *Eutrichomastix* especially by examining how the fourth flagellum was arranged. Hence, she suggested the placement of the newly discovered flagellate in a new genus which she named *Pseudotrichomonas* and therefore renamed the flagellate from *Trichomonas keilini* to *Pseudotrichomonas keilini* [26].

Prior to obtaining a true *P. keilini* culture and get a SSU rDNA sequence accordingly, the phylogenetic position of *P. keilini* was not resolved. It was classified to be part of the Honigbergiellida due to the absence of costa, comb-like structure and an infrakinetosomal body; also, because of the presence of three anterior flagella and an undulating membrane in a lamelliform [42]. One of the other reasons that gave support to that placement was that Honigbergiellida contained the only found free-living trichomonads.

Using a 53-taxon alignment, Yubuki and colleagues (2010) constructed phylogenetc trees for the newly discovered species in an attempt to resolve the parabasalian tree. Both Maximum Likelihood (ML) and Bayesian analyses produced the following strongly supported result: the Trichonymphida, Cristamonadida, Hypotrichomonadida and Spirotrichonymphida formed a clade. On the other hand, Honigbergiellida and Tritrichomonadida had a paraphyletic relationship and showed that the *P. keilini* strains belong to the Trichomonadida where they clustered instead of the Honigbergiellida [193].

Another interesting finding by Yubuki and colleagues (2010) was that the SSU rDNA genes of the two discovered *P. keilini* strains only differed in 21 base pairs out of 1,499bp [193].

Overall, *P. keilini* species did not branch with other free-living parabasalids like *Tetratrichomonas undula*. Free-living parabasalids generally did not group together in a clade, they rather formed lineages that were closely related to the Trichomonadida clade [193].

1.0.6.2 Habitat and living conditions

The discovery of *Pseudotrichomonas keilini* started in 1934 with a freshwater sample from a Lincolnshire pond that was then analyzed by Prof. Anne Bishop. In that sample, different species were identified like pink bacillus, but there was also many Protozoa in the sediment such as *Paramecium caudatum sp.*, *Cyclidium sp.*, along with some small sized amoeba, *Euglena virdis*, *Astasia sp.*, a *Hexamita* species in addition to some unidentified flagellates [24].

After several culturing attempts with different growing conditions, the flagellates grew best with a serum-saline medium with the addition of boiled wheat grain in hay infusion and in freshwater from pond or rain with a temperature of 4 to 31°C. It has also been detected that a mixed flora of bacteria is always present in the culture which later lead to the conclusion that *P. keilini* is feeding on the bacteria [24].

Following the detection of parasitic *Trichomonas* species outside a host, some species were described in drinking water [62] or a lettuce-stalk broth as in the case of *T. hominis*, speculations began to rise about whether or not *P. keilini* was a free-living one or was merely found in the pond with dysenteric stool [24]. Further investigations also confirmed that it was a free-living flagellate [25].

As *P. keilini* had a significant morphological resemblance with other trichomonads, there was an initial hypothesis that it was also parasitic. However, since the cultured cells died at 37°C suggesting that *P. keilini* does not infect warm-blooded animals. This led to another hypothesis that it could be a parasite for cold-blooded organisms since the amphibians-infecting *Trichomonas augusta* also died at 37°C. However, the ability of *P. keilini* to maintain a living under free-living conditions indicated that it is most likely to be a free-living protist [24]. The work of Yubuki and colleagues (2010) also confirmed the growth temperature for *P. keilini* as the isolated strains died at 37°C and grew at 16-24°C [193]

Although *P. keilini* was originally discovered in freshwater in Lincolnshire pond [26, 24] and was reported to be found in other geographical locations also in freshwater [18, 31, 113] , the newly isolated strains came from both marine and freshwater sediments. While *P. keilini* NY1070 (Japan, sampled from mangrove sediments was kept in seawater medium, *P. keilini* LIVADIAN (Cyprus) was collected from a pond with unknown salinity in Cyprus, but was maintained in freshwater medium. Both strains harbored identical morphology despite the changes in growth media and environments.

The distribution of costas across parabasalids can mean either that there were multiple independent origins of a costa, or multiple losses events for costa or a combination of both loss and gain events that happened independently [42]. While it has been more favoured previously that the costa had been through several losses events, the study done by Yubuki and colleagues (2010) shows that the loss of costa in *Pseudotrichomonas keilini* was a secondary loss [193].

Following all the previous analyses, the *Pseudotrichomonas* genus was moved to the order Trichomonadida in the family Trichomonadida instead of Honigbergiellida with the following taxonomy summary [193].

TAXONOMIC SUMMARY

Phylum Parabasalia
Class Trichomonadea
Order Trichomonadida
Family Trichomonadidae
Pseudotrichomonas Bishop, 1939

The phylogenetic analysis of data also shows that Parabasalia's sister lineage is Fornicata comprising diplomonads(*Spironucleus*,*Giardia*, *Enteromonas* and *Octomitus*) and retromonads (*Chilomastix*and *Retortamonas*)[77].

Although data shows that the most recent common ancestor of the Fornicata was a free-living one, it is still unclear how the common ancestor of parabasalids was like although we know that in their evolutionary history they evolved from free-living ancestors

We are yet to have the data to show whether free-living parabasalids reflect morphostasis as an ancestral mode of life or whether it rather reflects endobiotic ancestors. It is also possible that it could be a combination of both [193].

The number of discovered free-living parabasalids is only a tiny fraction of their total population on earth. The lack of data available about other organisms will still hinder a better understanding about the evolution of parabasalids.

1.0.6.3 Morphological features of *P. keilini*

Trichomonads often harbor several features that make them distinguishable. Such features include; an axostyle, an undulating membrane, a basal fibre,

One of the most distinguishable features of *Trichomonas* is the way they move. The same mechanism was detected in *P. keilini* with a cell size ranging from 7-11 x 3-6 μ , and sometimes only 5-7 μ. On average, *P. keilini's* structure is more rounded than that of *Trichomonas himinis* which forms a spike in the middle of the posterior end. Another distinct characteristic of *P. keilini* is its elasticity of form which was something different than the previously described *Trichomonas* species of that time.

They can be globular in some cultures, carrot-like shaped or extremely elongated in others. The age of the culture seems to be the main determinant for the shape that will be present. While younger cultures seem to have more of the globular form, older ones tend to contain more elongated types. However, regardless to the age of the culture, all of the aforementioned shapes were found. [24]

The initial morphological analysis of the unidentified flagellates showed significant similarities with trichomonads. They had three anterior flagella, but they had a shorter undulating membrane attached to the anterior end of the body. In addition, the bordering flagellum was also absent. [24]

The length of the flagella varied a lot. While they all arise from the anterior end, the longest is about almost twice the length of the whole body and the shortest is half the length of it. While the third one had a length that was slightly shorter than the longest. At the same point where the three flagella arise, an undulating membrane also arises. The flagellum attached to the undulating membrane does not project freely from its termination point. Moreover, a mouth that is well-developed can be seen at the anterior end in the same position as in other trichomonads. Several food vacuoles are also present. [24]

Examining fixed and stained samples of *P. keilini*, the flagella seem to rise from a basal body (formally knows as blepharoplast) that is well-developed. [24]

Despite resembling the main structure of other trichomonads like *Tritrichomonas fecali*, it differed in lacking a basal fibre which is is normally found in trichomonads below the undulating membrane. Another feature which also seemed to be frequently missing was the axostyle. [24]

Further investigation of the cellular structure showed that it is a uninucleate cell with a spherical or oval dense nucleus that can be found below the point where anterior flagella arise. It contains mostly a single karyosome, with two karosomes in some occasions. The axostyle was seen in some individuals while it was completely invisible in others as it is reduced and can only be seen with difficulty. In the case of the presence of axostyle it ends posteriorly in the posterior end of the body, particularly, in the sharp projection of it. However, in case of the absence of an axostyle, the sharp projection is even more noticeable. [24]

The two *Pseudotrichomonas keilini* isolates that were obtained from the work of Yubuki and colleagues (2010)[193] resembled the diagnostic features that were described in the discovered *P. keilini* by Bishop(1935, 1939) [24, 26]. Both species had an undulating membrane ,three anterior flagella, and an absent costa. In addition, different morphological features appeared to differ according to the growth phase. It was seen that dividing or immature cells only had two anterior flagella rather than three which is the number that is normally found.

Costas can be lost in several parabasalids such as *Simplicimonas*, *Monocer-*

comonas, the Dientamoebidae, the Spirotrichonymphida, the Trichonymphydia, the Honigbergiellida, and most cristamoads [42, 81].

Overall the cell size of the newly isolate strains ranged in size from 9.4μm in *P. keilini* LIVADIAN (Cyprus) and 10.7μm in *P. keilini* NY0170 (Japan) excluding the size of the axostyle.

While the undulating membrane was long enough to reach the posterior end of the cell, the posterior flagellum's distal end did not go beyond the undulating membrane. A large number of hydrogenosomes was also detected using the electron microscope.

1.0.6.4 Cell division and life cycle

During the several cultivation attempts, no cyst form was detected at all. Also, multiple fission that often occur in *Trichomonas* species [22] were never found. [24]

Cysts, comb-like structures or an infrakinetosomal body were also not observed in the *P. keilini* strains from Japan and Cyprus. In addition, the undulating membrane varied greatly in length, ranging from one half of the cell while in other cases it was long enough to reach the posterior end. Moreover, a parabasal body was observed forming a small darker disc near the nucleus [193].

Prior to cell division, Bishop detected that some granules appeared in the place of a single karyosome, in addition to an increase in the siderophic material. These granules then form five chromosomes. It is more difficult to count and clearly see the chromosomes in the case of cells which lack an axostyle, as the chromosomes are formed in a compact dense mass. A nuclear membrane cannot always be detected at the time of cell division due to its thinness. In case of a visible nuclear membrane, spindle fibres were not detected. [24]

The origin of axostyles seems to form a mystery. While they were proved to be originated during the telophase from daughter blepharoplasts in trichomonads [22, 23]. The method of division is generally similar to that of *T. hominis*, *T. sanguisugae* and *T. batrachorum*. [24]

During the prophase, the number of chromosomes in *T. sanguisugae* and *T. batrachorum* is six while only five chromosomes were detected in *P. keilini*. Moreover, another difference in that a basal fibre is missing in *P. keilini*. Another origination mystery is that of the undulating membrane along with its flagellum. It was unknown how they arise, however, is some samples, prior to nuclear division, a flagella and two membranes can be seen. One unusual phenomenon that happens in later stages of division was the free projection of the flagellum attached to the undulating membrane beyond the body's edge. [24]

Here, we present the draft transcriptome and proteome of *Pseudotrichomonas keilini*, the free-living relative of the parasite *Trichomonas vaginalis*.

1.0.7 Project objectives

The motivation to conduct this project was due to the limited data available for free-living parabasalids, with most of the data describing the parasitic ones. This has caused a conundrum in the complete understanding of the evolution of parabasalids and the early lifestyle of their last common ancestor. Through studying the first free-living parabasalid discovered, *Pseudotrichomonas keilini* we aim to fulfill the following targets:

1. Sequence, assemble, and annotate the genome and transcriptome of *P. keilini*.
2. Understand the genomic and transcriptomic structures of *P. keilini*.
3. Discover the metabolic capabilities of *P. keilini* as a model organism for free-living parabasalids.
   (a) Identify the organelle used for energy metabolism in *P. keilini*.
   (b) Highlight the primary pathway used for energy production.
   (c) Draw a map of the different metabolic pathways present in *P. keilini*.
4. Uncover the phylogenetic position of *P. keilini* among other relatives.
   (a) Produce a species tree for *P. keilini* and other parabasalids.
   (b) Find eukaryotic marker proteins in *P. keilini* in comparison to other eukaryotes.
5. Genomically compare *P. keilini* to its parasitic close relatives; *Trichomonas vaginalis* and *Tritrichomonas foetus*.

Chapter 2

Methods

2.0.1 Cell Culture

Pseudotrichomonas keilini cells were obtained from the deposited NY0170 (Japan) strain which was deposited at the American Type Culture Collection (ATCC) under the accession number PRA-328. Cells were cultured in the same way described by Yubuki and colleagues (2010) using an ATCC medium 1034 which is 5% modified PYNFH medium for the enrichment and maintenance of the *P. keilini* culture. To isolate a single cell, we used micropipetting of an enrichment culture. The clonal culture was maintained and established at 22 °C with weekly passages[193].

This part of the analysis which includes cell cultre, generation of RNA and DNA material was done by our collaborators at the Arizona State University (ASU) in the lab of Dr. Gillian Gile.

2.0.2 Assembly

2.0.2.1 Transcriptomic data sequencing (RNA-Seq)

Transcriptomic data was sequenced on four runs by Génome Québec in Canada, each with different read length. The first three runs were sequenced using Illumina's MiSeq technology with a variation in the number of reads. However, the fourth run had a significantly larger coverage as it was sequenced using Illumina NovaSeq 6000 S2 PE100 - 50M reads.

2.0.2.2 Assembly of individual runs

Different approaches and software in assembling each run on its own were used to in order to come up with the best pipeline. The first thing we wanted to test was if trimming the reads using trimmomatic-0.38 [28] can obtain a better result. As we did so, the result was as follows for the second run: 14,539 transcripts compared to

28,595 transcripts with the paired and unpaired trimming, we also assembled the data directly without trimming which resulted in 41,897 transcripts. That showed a dramatic decrease in the number of reads due to trimming which was worrying, that we could be losing some of the *P. keilini* reads. The second step was testing the best software to assemble the reads. We used Trinity-v2.8.4 [76] and rnaSPAdes v3.12.0 [36] and tested the assembly quality using QUAST-v5.0.2 [75]. We obtained the following data for Trinity-v2.8.4 and rnaSPAdes v3.12.0 in the assembly of the first run, respectively: the number of transcripts was 23,989 and 8,220. The total length was 5,923,855 bp and 5,039,631 bp. As for the N50, the values differed with 836 kb compared to 698 kb. Which means that half of the sequence is in contigs larger than or equal to 836 kb or 698 kb. Based on these results, we decided to continue assembling the reads without trimming and using Trinity-v2.8.4 for the assembly.

The assembly of the next three runs resulted in the following number of transcripts: 41,897, 70,914 and 73,607 respectively with the corresponding quality assessment results of total length and N50 value: (43,935,258 bp and 3,647 kb) , (45,485,967 bp and 1,812 kb) and (37,133,889 bp , 1,681 kb).

2.0.2.3 The assembly of combined runs

Following the same assembly methodology mentioned in 2.0.2.2, we combined the reads of the different runs together and assembled them to get the highest coverage of data. The first and second sequencing runs were combined into 19,940,827 reads and then assembled into 55,051 transcripts. After obtaining the data from the third run, assembling the first three runs with 28,697,690 reads resulted in 92,976 transcripts. The final and last concatenation was done subsequent to acquiring the fourth sequencing run of transcriptomic data, which resulted in 137,389 transcripts from assembling 90,695,489 reads.

2.0.3 Structural Annotation of Assembled Transcripts

Protein prediction for transcripts was done using TransDecoder-v5.5.0 [76] using the following steps: (i) finding the long Open Reading Frames (ORFs) which have a minimum length of 100 amino acids using TransDecoder.LongOrfs. (ii) predicting the coding regions of theses ORFs with TransDecoder.Predict. The number of proteins predicted from each run varied and they were 10,475 , 50,296 , 46,200 and 34,498 accordingly.

As for the result of the combined asssembly, it was as follows: for the first and second sequencing runs it was 56,345 while for the first, second and third sequencing runs the number of predicted proteins was higher with 83,878 proteins in total. Lastly, the third and final combined assembly combining the four runs generated

83,266 proteins.

2.0.3.1 Contamination filtering of transcripts

The main challenge for this project was to obtain a set of eukaryotic transcripts without bacterial contaminants. That is due to the nature of *Pseudotrichomonas keilini* being a bacteriovore that does not grow in pure culture, so our transcriptome assemblies contained many transcripts from contaminating bacteria. We applied a variety of approaches to identify the subset of transcripts that likely originated from *P. keilini*, and this was a major element of the analysis work of my project. One of the attempts to clean the transcriptomic data was to use a tool for fast functional annotations of sequences which was done for the first three assembled runs. By uploading the predicted proteins dataset to eggNOG mapper-v 4.5.1 [94] and setting the default parameters, we only found 4,636 eukaryotic proteins. Hence, we used another method in which we were aiming to find proteins that have high similarity match with the proteomes of the close well-studied relatives *Trichomonas vaginalis* and *Tritrichomonas foetus*. We then used BLAST [7] to find the proteins that have their first hits with these organisms from the nr database in National Center for Biotechnology Information (NCBI). Followed by Entrez Programming Utilities [161] and a homemade Python script implementing Pandas library [130, 183] to filter-out the result. This approach resulted in only 5,027 filtered proteins in total. So we needed to apply a different procedure.

After this period of experimentation, we settled on the following protocol, which resulted in a final set of 19,221 transcripts. According to BUSCO, [168] which uses the conserved single-copy genes across eukaryotes as a reference for the completeness of the data *P. keilini* contains 20 complete BUSCOs compared to *Trichomonas vaginalis* which has 22 BUSCOs. That means that the set of transcripts obtained likely represent most of the protein-coding genes present on the *P. keilini* genome.

This contaminant filtration process was done after the concatenated assembly of the four RNA-seq runs and their protein prediction which resulted in 83,266 predicted proteins in total. We then ran a Diamond BLAST search for all the aforementioned predicted proteins against a custom database containing all predicted proteins from 38 published excavate genomes including parabasalids, along with a representative sampling of 73 other eukaryotes, 148 bacteria and 146 archaea (see Supplementary Table A1). For any of the queried proteins to be considered a putative *P. keilini* one it had to fulfill either of the following criteria (i) to have a best hit with a parabasalid relative of *P. keilini*, or (ii) alternatively, its first three Diamond BLAST hits were eukaryotic in the custom database [33].

Although this approach is being very robust to avoid any false positives of probable contaminants, it has the downside of being over-conservative and we may have missed proteins

This resulted in 21,300 putative *P. keilini* proteins.

2.0.3.1.1 Similarity-based protein clustering

As Trinity sometimes predicts closely similar partial proteins from different transcripts, we clustered proteins that were 100% identical over the overlapping length using CD-HIT 4.8.1 [119] the result was a total of 18,851 clustered proteins.

2.0.3.1.2 Testing the transcriptome completeness of *P. keilini*

As we followed a strict contaminant filtering pipeline to remove contaminants from the transcriptomic dataset, we needed to make sure that did not affect the completeness of the *P. keilini* proteome size obtained from this process.

In order to do so, we assessed the *P. keilini* proteome completeness using Benchmarking Universal Single Copy Orthologs BUSCO [168] on a dataset of 33 species of metamonads. BUSCO works by searching for the conserved gene orthologs among the selected species in the query against thae dataset. In our case the query was the proteomes of both *P. keilini* and *T. vaginalis* while the dataset was that of the selected species of metamonads.

2.0.4 Functional annotation of transcriptomic data

In order to understand the metabolic capabilities of *P. keilini*, we tried different tools to annotate the function of the predicted *bona fide* protein set.

We first started by assigning functions to the predicted proteins using one of the tools in Kyoto Encyclopedia of Genes and Genomes(KEGG) which is called blastKOALA. The idea of how blastKOALA works is that it is an annotation server for genomic sequences which annotates queried sequences by assigning KEGG orthology (KO) numbers to them. Each of these KO numbers represents a functional ortholog for proteins and genes based on the data present from biochemically-annotated proteins according to scientific papers. This data which shows the biochemical interactions and functions of different proteins is stored in the PATHWAY database.

KEGG database provides a computerized source for the current knowledge on protein functions in different metabolic pathways [99, 100].

Out of the total 18,851 proteins, only 11,674 proteins were annotated while 5,204 of these proteins had a first-hit with a KO number which were the most credible according to our quality tests.

Our quality tests involved doing BLAST[7] searches for some of the results that seemed controversial, for example if we found a complete or nearly complete Krebs cycle that would require further investigations since its close relatives *T. vaginalis,* and *T. foetus* have incomplete Krebs cycle set of enzymes. After following that test it appeared that the result of a KO first-hit is the most credible in the annotation of KEGG database. That leaves only 5,204 proteins.

That number of annotated proteins was not enough compared to the expected proteome size based on the available data for close relatives, so we decided to try other tools for the annotation.

The second tool we used applies orthology assignment based on the eggNOG database which contains phylogenies and previously-computed clusters that are used to functionally annotate proteins [93, 95]. Using eggNOG-mapper v2 online tool, which uses eggNOG 5.0 clusters and phylogenies that is a more expanded database which covers a larger number of genomic sequences. We were able to annotate a total of 12,712 proteins out of the original 18,851. We chose to annotate the proteins with eggNOG mapper because it links several databases together, one of which is the KEGG database. Out of the 12,712 only 8,059 had a KO assignment with eggNOG. To test the assignment and annotation quality of eggNOG mapper, we also applied the same quality test we used for KEGG's blastKOALA results to determine the accuracy of eggNOG mapper's annotation. The annotation passed our quality check which left a total of 6,139 proteins with no function assignment or annotation by eggNOG mapper. In an attempt to annotate these proteins, we first checked if they had a KO assignment with a first hit from the previous annotation with blastKOALA. We found a total of 99 proteins that were annotated by blastKOALA but not eggNOG mapper and with KO assignment. Thus, the total of annotated proteins is 12,811.

2.0.5 Gene finding

2.0.5.1 Genomic data assembly

Genomic data was sequenced using Single Molecule, Real-Time (SMRT) technology developed by PacificBiosciences (PacBio) also by Génome Québec in Canada. The SMRT technology allows faster sequencing with higher coverage [55]. The data was

sequenced with 100x coverage Sequel technology, using Sheared large insert library type. We used the Filtered Subread Bam files containing a total of 3,643,047 reads.

Our initial hypothesis for the genome size of *P. keilini* was that it is will be similar to that of its close relative *T. vaginalis*. The genome of *T. vaginalis* was discovered to be composed of highly repetitive regions with an estimated size of nearly 160 mega bases.

2.0.5.2 Contamination filtering of DNA reads

To obtain a filtered set of *P. keilni* reads, we tried different approaches using different software programs for PacBio reads assembly and error correction and they were later assessed based on contig length and BUSCO completeness [168].

The first approach was done with Flye [106] which works on raw PacBio reads and produces polished contigs. We sat the parameters with an expected genome size of 160 mega bases according to the size of its close relative *T. vaginalis*. That produced 80 contigs which are longer than or equal to 1000 bp and 19 ones longer than or equal to 50,000 bp. Flye v2.6 output was later assessed using BUSCO 4.0.2 against a reference database of eukaryotes eukaryota_odb10 with 70 species and 255 BUSCOs. That resulted in only 33 Complete BUSCOs (C), 21 Complete and single-copy BUSCOs (S), 12 Complete and duplicated BUSCOs (D), and 1 Fragmented BUSCOs (F).

Another approach was done in which we tried another software for reads assembly and polishing. Canu v1.9 [107] was used with the same parameters and the same quality assessment test was conducted using BUSCO 4.0.2 which resulted in 32 Complete BUSCOs (C), 20 Complete and single-copy BUSCOs (S), 12 Complete and duplicated BUSCOs (D), and 1 Fragmented BUSCOs (F). As for the size of contigs, Canu produced 91 contigs which are longer than or equal to 1000 bp and 34 ones longer than or equal to 50,000 bp.

Since the BSUCO completeness result was highly similar in both assemblers, we relied on the contig size as a quality reference and hence, we decided to continue the analysis using Canu's output.

2.0.6 Metabolism of *Pseudotrichomonas keilini*

2.0.6.1 Detection of the energy production organelle

Falling in the excavates group where different kinds of mitochondrial-related organelles (MROs) are present, it was expected to find an MRO in *P. keilini*. However, the MRO type in *P. keilini* was unclear. Being a close relative of two well-studied

hydrogenosome-containing species; *T. vaginalis* and *T. foetus*, we started by trying to identify key enzymes in different MROs including the hydrogenosome. .

We used the supplementary table in Stairs et al. (2015) [171] which included key enzymes of each of the five classes of mitochondria and its related organelles classified by Muller and colleagues [136]. Using Diamond [33] we blasted the sequences against the database of *P. keilini* filtered proteins and we identified key enzymes. Other protein sequences were added for these enzymes by doing a BLAST search against nr database and selecting five sequences for Eukaryotes, Bacteria, Archaea , and most importantly, Alphaproteobacteria.

The obtained sequences were then aligned using mafft v7.390 [104]. After checking the alignment, and selecting the most complete sequences, we then ran BMGE-1.12 [44] which trims the multiple sequence alignments and produces phylogenetically-informative regions.

The aligned selected sequences from BMGE were then used to construct a maximum-likelihood tree with IQ-TREE-v1.6.10 [141]. After running a model test analysis, the most appropriate model chosen was LG+C60. That model was then used for the construction of phylogenetic trees for hydrogenosomal proteins.

2.0.6.1.1 Complexity of the *P. keilini* Hydrogenosome

From our previous analysis, we detected most of the key hydrogenosomal enzymes in *P. keilini* that were identified in *T. vaginalis*. Following that, we attempted to unveil how many of the 569 proteins that were found in the proteome of *T. vaginalis* hydrogenosome [164] can be detected in the hydrogenosome of *P. keilini*. Since we are working with protein sequences, we retrieved the protein sequences using the gene IDs that were provided in the supplementary material in the Scneider et al.(2011) [164] using Batch Entrez's command line tool [187] and we found 487 proteins.

One of the ways to assess the complexity of the *P. keilini* hydrogenosome was to find the number of the Mitochondrial Carrier Family (MCF) proteins. These proteins are found on the mitochondrial membranes and are responsible for the transfer of molecules across the membranes [144]. We did that by doing an hmmsearch with HMMER 3.2.1 (http://hmmer.org/) . We used the PF00153 hmm profile for the MCF proteins which contains 125,808 sequences from 1,117 species including *Trichomonas vaginalis*. HMMER profiles are probabilistic profile hidden Markov models [51, 53, 110] which are used to accurately find distant protein homologs. By aligning the *P. keilini* proteome against the hmm profile, with a threshold value of .00001 for the e-value, a total of six proteins from *P. keilini* was aligned. Given that a close number of five proteins was aligned in *T. vaginalis*, the initial thought was that the hydrogenosomes of both *P. keilini* and *T. vaginalis* are highly similar .

2.0.6.1.2 Hydrogenosomal membrane transporters

One of the key things we needed to understand after the detection of the hydrogenosome in *P. keilini* is the type of inner and outer membrane transporters that are found on the surface of the hydrogenosome.

To determine the presence/absence of such proteins, we made a homemade script with the following pipeline. i) Do a protein BLAST search [7] for each of the queried protein against a database of parabasalids including *P. keilini*, other metamonads, and eight bacterial representatives. ii) Make a list of the absent or present proteins in each of the species based on the BLAST search. iii) Align the protein sequences using mafft v7.390 [104] and build phylogenetic trees for each of these alignments using IQ-TREE-v1.6.10 [141] by finding the best-fit model. iv) Based on the trees topology, we could confirm the presence or absence of the queried proteins by discarding any *P. keilini* proteins that were grouping with bacteria.

2.0.6.2 Lipid metabolism in *P. keilini*

One of the methods we attempted to discover the preferred metabolic pathway of *P. keilini* was to compare it against those of its close relatives. One of its few free-living closely-related organisms is the flagellate *Naegleria gruberi*, which prefers lipids as a substrate that is then metabolised for energy production [20]. Also, since *P. keilini* feeds on bacteria, that made the lipid metabolism argument even stronger, given the biochemical composition of the bacterial membranes of phospholipid bilayer. We conducted this analysis by doing a BLAST search [7] against the enzymes associated with lipid metabolism in Table S2 of Bexkens et al. (2018) against the curated protein set of *P. keilini* followed by the construction of phylogenetic trees which then determines the presence or absence of the queried enzymes. The method is explained in detail in 2.0.6.1.2.

2.0.6.3 Phylogenetic trees of conserved eukaryotic proteins

Williams and colleagues (2017) identified 44 conserved proteins across eukaryotes [189]. We used the HMM alignment profiles for these proteins to do an HMM search against the proteome of *P. keilini* and we identified 41 proteins that matched the threshold e-value score of the alignment(0.00001). In addition, we constructed phylogenetic trees for these proteins following the method in 2.0.6.1 using mafft and BMGE for alignment [104][44]. We then used Q-TREE-v1.6.10 [142](see Supplementary Figures 3.17 to 3.57).

Chapter 3

Results and Discussion

3.0.1 Genomic and transcriptomic sampling of the *P. keilini* genome

As we were unable to obtain a pure culture of *P. keilini*, we performed deep sequencing of mRNAs and DNA followed by a multi-step filtering process to identify and remove reads, assembled transcripts, and genome contigs that were not derived from the *P. keilini* target. We performed four rounds of transcriptome sequencing using Illumina MiSeq instrument using version 2 chemistry, 2x250 paired end reads and NovaSeq 6000 S2 PE100 - 50M reads, yielding a total of 90,695,489 reads. 137,389 transcripts were assembled using Trinity-v2.8.4 RNA-Seq [76], proteins were predicted using TransDecoder-v5.5.0 [76] and then searched against a database of the complete coding sets for a selection of other eukaryotes and complete bacterial and archaeal genomes (Table A1). We classified transcripts as *bona fide* eukaryotic proteins if either (i) the best hit was from a parabasalid relative of *P. keilini*, or (ii) alternatively where the first three database hits were eukaryotic in a Diamond BLAST search [33] were retained as putative *P. keilini* proteins that resulted in a final set of 19,221 transcripts and 21,300 proteins which were further reduced by similarity using CD-HIT 4.8.1 [119] to 18,851 encoded unique proteins that likely derive from *P. keilini*. We predict that this transcript set captures most of the protein-coding gene content of *P. keilini* because BUSCO [168], which uses the conserved single-copy genes across eukaryotes as a reference for the completeness of the data, showed that *P. keilini* contains 20 complete BUSCOs compared to *Trichomonas vaginalis* which has 22 BUSCOs. Another reference for completeness is that the number of proteins is fairly comparable to its close relative *Tritrichomonas foetus* which has 23 thousand proteins.

After finding the *bona fide* proteins, we assigned functions to them by orthology assignment based on the eggNOG database which contains phylogenies and previously-computed clusters that are used to functionally annotate proteins [95,

93, 94]. Using eggNOG-mapper v2 online tool, which uses eggNOG 5.0 clusters and phylogenies that is a more expanded database that covers a larger number of genomic sequences. We were able to annotate a total of 12,712 proteins out of the original 18,851. We further used KEGG's blastKOALA to annotate the rest of the unannotated proteins and annotated 99 more proteins. Thus, the total of annotated proteins is 12,811.

3.0.1.1 Genomic data assembly

To reach a filtered set of *P. keilini* contigs, we assembled the raw PacBio reads using Canu v1.9 [107] and sat the parameters to an expected genome size of 160 mega base which is based on the genome size of *T. vaginalis*. That resulted in 189 unitigs and 91 contigs with 34 contigs and 76 unitigs of them longer than or equal to 50,000 bp. The total size of contigs was 43,993,368 bp with 54.15% GC content and 47,025,226 bp and 53.70% for the unitigs. BUSCO 4.0.2 [168] analysis was conducted to assess the genome completeness using a database of 70 eukaryotic species with 255 Total BUSCO groups searched and the following result was obtained:

- 32 Complete BUSCOs (C)
- 20 Complete and single-copy BUSCOs (S)
- 12 Complete and duplicated BUSCOs (D)
- 1 Fragmented BUSCOs (F)
- 222 Missing BUSCOs (M)

3.0.2 The hydrogenosome of *P. keilini* and reductive evolution of mitochondria in Parabasalids

The detection of hydrogenosome was done on two levels; a macro one which involved detecting the surface and structural proteins that were identified in *T. vaginalis* and a micro one for the key enzymes that are found in hydrogenosomal-containing organisms.

3.0.2.1 Identification of hydrogenosomal surface proteins

According to Schneider and colleagues in 2011, 569 proteins were identified from the purification of the hydrogenosomes of *Trichomonas vaginalis*. This number indicates that the hydogenosome's proteome is almost half of the mitochondrial one which ranges from 1,000–1,500 proteins.

Based on their data, 123 proteins were found on the surface of the hydrogenosome of *Trichomonas vaginalis*. Using the method specified in 2.0.6.1, we identified 120 proteins out of them which can be seen in Table A2

Moreover, 413 putative membrane proteins, enzymes and hypothetical proteins were identified in the hydrogenosome of *Trichomonas vaginalis* , and we detected 333 proteins out of them in *P. keilini* (see Table A3)

Schneider et al.(2011) also identified thirty-three proteins as probable contaminants for they are homologues of proteins found in the organelles of other organisms such as the nucleus, endoplasmic reticulum, or vesicles. Through our analysis, we found 32 of these proteins (see Table A4). In their paper, they proposed further analysis for these proteins due to the difficulty in purely isolating the hydrogenosomes [164].

3.0.2.2 Identification of key hydrogenosomal enzymes

The second level of detection was done by following the method described in 2.0.6.1, key hydrogenosomal enzymes were detected in *P. keilini* that were identified in the close relative *T. vaginalis*. One of these key enzymes is pyruvate: ferredoxin oxidoreductase (PFO) which is responsible for the oxidation of pyruvate to produce acetyl-Co-A and CO_2 [171].

In addition to 12 other key enzymes such as [FeFe]-hydrogenase (hydA) and its three maturase enzymes; small GTP-binding protein (hydF), FeFe-hydrogenase assembly protein (hydG) and radical SAM domain containing protein (hydE). We also identified acetyl-Coenzyme A and its B subunit (ACST1B), in addition to three of the Glycine cleavage system enzymes; the H-protein (GCSH), L-protein (GCSL), and Serine hydroxymethyltransferase (SHMT). As well as the enzymes present in the Electron Transport Chain complex I; NADH-quinone oxidoreductase subunit F (NUOF) and NADH-quinone oxidoreductase subunit E (NUOE). As for the Krebs cycle enzymes, we detected two of them; Succinyl coenzyme A synthetase (SCS) and malate dehydrogenase.

Table 3.1: List of key enzymes detected from the metabolic pathway of the *P. keilini* hydrogenosome

Enzyme Name	Sequence in *P. keilini*	Accession number basted against	Accession number Organism
PFO pyruvate: ferredoxin oxidoreductase	>Pseudotrichomonas_keilini_TRINITY_DN39_c0_g1_i13.p1_PFO MGELMDAWIAQGRKNIWGNPVKLAMLQAEGGAAGAVHGATTVGTLCTTFTASQGLLLMIP DIYKIAGEYCPAVFHITARSIAQSALSIYNDHGDIYAARSTGIPMLCSNGVQEAHDMAAI SHLTTIKTGLPMMHFFDGFRTSHEINTYEEIDNEVLAKMIDPKGLARIRARALNPEHPKV QGTICGPEYVWQTVEKLRKPYQLLPREIEAMMVEFGKNTGRVYKPYQYVGPMNAENLIIM MGSGCDPVEEYVVTHPQSSIGILKIHMLRPFSIEMFNQAIPQSVKKICVLDKYSDPTGAR EPLFTDVATAVTNKRNVQIIGGRYGISSRDFAPPHVDAIVKNLLQPNSLDGFTVGVNSPE TALPIGIPFDNLPQTTRQCIFWGLGSDGTVGANKEAIKLIVDNTEMYGQAYFAYSAHKSG GLTTSHLRFGEQPINAPYFIQSADYIACHNPAYLYKFDMLKPLKQGGVFVINISSKTNLD TDLPSSVRRTLAEKDAKLYTIDATQLAIDLGLPGRINMIMQTVFFGLSGVLPSVQCIELL KKSIKKQYIRKGQEVIQKNWDMVDAALAGLKEIKFDKNTWALAIDPPQPEKKGMARILDM TIKQLGEDVSVEEMTEIAQCPPGTAKFEKRGIAVNVPIWDDKKCIQCNTCSVLCPHAVIR PFLLTAEEAKGMKTQKAKGKEIKDYLYRIQISPLDCTGCGTCAAACPVQALTMTHRTPVH DETEGKNFETCMNVPNRGNLVNKFTLRGSQFQQPLLEFNGACPGCGEPAIIKLITQLYGD QLYIANATGCSLIWGATFPWNPYTVNEKGHGPAWANSLFEDNAEFGFGMFHSIEARRNIA KNLIVDLLKDGEFKGELKQKMEAILPVWNEQDSSAKLAEEIKPLLAAIQNPTEKMINLQS QADVLAQKSVWIIGGDGWAYDIGYGGVDHVLASGDNIKIVVMDTEVYSNTGGQCSKATSR GAVANFAAAGYAKGKKDLGSIAMTYGNIYVASTCLLADPAQAFKAITEAKEYNGPALIIN YSPCINHGVKKGMGSGPNHCKDLVKSGYVSLYRYDPRKVAQGKNGFVLDSSEPTYEIESL LKDENRYAALQDIYPTEAQTKYPALVEDLKKRYNIYKALASKQ*	XP_001582360	*Trichomonas vaginalis*
HYDE radical SAM domain containing protein	>Pseudotrichomonas_keilini_4runs_TRINITY_DN77_c0_g1_i1.p1_HydE MLYTLHHAKRELPLFETALKNALSGFKMTHDEIVTLLQANQPDQIAQLHEAAGNVTQKVF GNEVSIRGIVEFSNHCEKNCHYCGVTAYEDKFLIPHESILECCDFMWSKGYRNIVLQSGE VTSKQRIDWVANLLEKIFNKFGKEKDTGMCVILSIGELSYEQYKRLYDIGAQRFLLRIES SNPKLYASIHPPDHLYERRIQCLKDLKSIGYVAGTGSMVGLPGQTYDDIAADIEFFRDNQ YPMIGLGPYIIHKDTHMGRELFKITTEDERKEADILKVQATLNVYDTIRLACPLNNIAAT TALDTLSPGAKAIALRGGSNVVMPIITPKLFRSGYQLYEGKKEVDEDREQTHQRVLNLMQ QIKKVPIFNRWNHPPLFTKLQNKNE*	XP_001326754	*Trichomonas vaginalis*
HYDA 64kDa iron hydrogenase	>Pseudotrichomonas_keilini_4runs_TRINITY_DN13155_c0_g4_i1.p1 MLSSVTISQSNVFNSLVRNISVNVNGKILEAKKGETILQLCDRNNIKIPRLCYHPNLPPK ASCRVCLVECDGKWLAPSCVTEVYDGLKVETKSPKVKSSVSNNLKLLLASHDETCSSCVA NHRCEFRDFVVQAGVTCPKRENPAPEKIDRSTNSIQIDTSKCVLCGRCVSACDSIAGQSA IIFGNRAKNMTVQPSTGLTLQDTSCIKCGQCTLYCPVGAITEKSQVLEVMDRLATKKHKL SVVQVAPAVRVAISEALGLPIGTNSQGKLVSVLRAIGFDLVYDTNYSADLTIVEEANELI HRLKDPNAVLPMFTSCCPAWVNYVEQSRPEFIPNLSSCRSPQGMLSSLIRNYLPKLNGIK PEDVFSVSIMPCTAKKDEIERPELKNSDGIKETDYVLTIRELVEIIKLSGIQYSSLPDSE FDTLFGVGTGAGQIFAATGGVMEAAARTAFEAITGKKLTKVEITSVRGMEESVKIAELDL DGTKLKVAVVHGIAPTSKLLDRLQSGDPELKDIKFIEIMACPGGCVCGGGTPQPKSKDVM SSRLSSIYRIDDLSKLRKSHENPMVIELYDKLLEKPNSHLAHELLHTHYKPHPKN*	XP_001305709	*Trichomonas vaginalis*
HYDF small GTP-binding protein	>Pseudotrichomonas_keilini_4runs_TRINITY_DN350_c0_g1_i2.p1_HydF MISYFKRYFASAPDLPRTHISLVGFMNAGKSTVMNAITQQPTSIVDSTPGTTADTKISLM EIHSIGPCKLIDTAGIDEMGNLGKKKLMKTISAVKESDVVLMIIDPTNFRIEPIELSDI AIRREKQVALVFNHFHDKEVDFEKLTNDIAKKLEKVCNKHLPKLNVSAINQKDANKAIVN FISTIKKQKQATPVIPPSYVGPHKNIFLNIPLDVESPSGRLLRPQTMVIEYSLRNQSSVF CYNMDLVSARGSKDKSTEEQKRFLEAINRSSPSLVITDSQAIDVMSKWTPINFPLTTFSV AMANFQTDGGLKQFLRGIDSLSKLKPGDKVLICEACNHDRIGDDIGTVQIPTKLKSIYPK VELDWSFGRAYEQKNLKEYALALHCGGCMISKQQMVSRLQDLFETGVPVANYGLALSWLA SPKALERVLQPWQ*	XP_001309182	*Trichomonas vaginalis*
HYDG Fe-hydrogenase assembly protein	>Pseudotrichomonas_keilini_4runs_TRINITY_DN272_c2_g3_i1.p1_HYDg MIGHHLSNYSSSFFRSISKQCITGTKWSPEREKYELPDPRQIVREDEINESLVLSRDSGR DVGAVRAILQKARERATMKDVPASNKEFMLGVDIDEAATLLNVPTENSSLIDEILDTAFY IKQKIYGNRIVLFAPLYVSNYCDCSCQYCGYRGTNTTIQRTKLSDDQVKAEVETLEKMGH KRILMLTGESPEYTFEDFLGALKAASSVKTGKSGEIRRINVEIPPLSVTDIKRLKGVGCV GTFTVFQETYHRKSYAKYHPYGPKADYDYRITCMDRSQMGGIDDVGMGALFGLYDHKFEV LALLQHAQHLDRTYGTGPHTISFPRIRPATGTPLSEHPPHCVDDLDFKRLIAVMRCAVPY TGMIISTRESIKMRNELLKIGISQLSAASSTEVGSYTQSQTDGSRGQFSVFDHRPVDDVV SGLMRDGYVPSWCTACYRLGRTGETFMKWAKSGEIHRNCHPNALFTLAEYLMDYAPEQTK KTGWNLIENEIEKIDDIKRRNQTKERIQLIKGGKRDLYF*	XP_001313153	*Trichomonas vaginalis*
Fe-Fe hydrogenase	>Pseudotrichomonas_keilini_4runs_TRINITY_DN13155_c0_g4_i1.p1_Fe-Fe-hydrogenase MLSSVTISQSNVFNSLVRNISVNVNGKILEAKKGETILQLCDRNNIKIPRLCYHPNLPPKASCRVCLVEC DGKWLAPSCVTEVYDGLKVETKSPKVKSSVSNNLKLLLASHDETCSSCVANHRCEFRDFVVQAGVTCPKR ENPAPEKIDRSTNSIQIDTSKCVLCGRCVSACDSIAGQSAIIFGNRAKNMTVQPSTGLTLQDTSCIKCGQ CTLYCPVGAITEKSQVLEVMDRLATKKHKLSVVQVAPAVRVAISEALGLPIGTNSQGKLVSVLRAIGFDL VYDTNYSADLTIVEEANELIHRLKDPNAVLPMFTSCCPAWVNYVEQSRPEFIPNLSSCRSPQGMLSSLIR NYLPKLNGIKPEDVFSVSIMPCTAKKDEIERPELKNSDGIKETDYVLTIRELVEIIKLSGIQYSSLPDSE FDTLFGVGTGAGQIFAATGGVMEAAARTAFEAITGKKLTKVEITSVRGMEESVKIAELDLDGTKLKVAVV HGIAPTSKLLDRLQSGDPELKDIKFIEIMACPGGCVCGGGTPQPKSKDVMSSRLSSIYRIDDLSKLRKSH ENPMVIELYDKLLEKPNSHLAHELLHTHYKPHPKN*	XP_001305709	*Trichomonas vaginalis*
ASCT1B acetylCoA hydrolase/transferase subtype 1B	>Pseudotrichomonas_keilini_4runs_TRINITY_DN514_c0_g2_i1.p1-ASCT1B FIFNFESFDSHNIMLANVFNSRGFIRRNRNPIPYKGFNPIPGKNPKECNTALEAVQCIKS NDRVFITEVAASPHELLKALDLRHKELQNVELTGIFPTGIEPLLVDDEAHNRAFIANCHF IGPPSRKSVLRKNPHHCFIPMFLHEIPQHLSSPDFPIDVALVSLSPPDKDGFCSIGPSVC CGRAGTDAAKIIVAEINPNIPYTLGNTKVHWSHLDYVFHTNRQIPQFPLQKITPEQASIG KYIAELVPDGACLQAGYGGVPDAVLNALKNHKHLGVHTEMFAEGLIDLIKSGAVDNSQKS YYPGVVSVSFVTGTDRLYKEISNNPLYHFEPVNVTNDPHNIAKNKNVISINSALQVDLSG QICADSMGPLQFSGVGGQVDFVRGASLSEGGRTFICLPSTASRGQVSRISPKLTHGSSVT TARWHGPTIVTEFGIAELWGLNTRQRAAALINIAHPKFREQLAREAAELYGTQ*	XP_002674540	*Naegleria gruberi*
GCSL glycine cleavage system protein L	>Pseudotrichomonas_keilini_4runs_TRINITY_DN218827_c0_g1_i1.p1_GCSL MLNCIKRAFTTTPDLLIIGGGPGGYAATIRAAKLGIKTVCIEKEELLGGTCLREGCIPSK YLLNLSHKVHEANHEFRNLGIKLPGKVEVDMVATQKRKNAVVTGLSKGIEYLIKKAGGER INGTAFIHSKDDIEVKQKDGKSIHFSPKNLLIAAGSNIWYPSSFPVDEKTIVTSRGALEM KEIPKSLCVIGGGVIGLELGSVWSALGSKVTIVDMASRVGGPSLEPEASQLITRVLKKRG MDFVLGKGVDSLKKSEQGVEVVVGDKTILAEKALVAIGRSANLKGYGLENLKLAMTERGL IKVDHKLATSVPNVYAIGDIVPGPQLAHKAEEEGIACVEHLAGHESSYNPDIIPSVIYTS PEIATVGITKEEAIKRGIQPIVKTVPYQSNSRARAILETEGFCTFVCDPKGTILGMHIVG PNAGEAIMEGTIAMKNNLKIDSIADTCHPHPTLSEAIMEAAKSILGKSVNF*	XP_001330332	*Trichomonas vaginalis*
GCSH - Glycine cleavage system H protein	>Pseudotrichomonas_keilini_4runs_TRINITY_DN183618_c0_g1_i1.p1_GCSH MISTYFNIRNYAKFFAPSHEWIDIEGKIGKIGISSFAEHQLGEVNIVDIQVGKIVKKEEE FGNIEAAKATSPIMAPMSGKILEINPAVQQTPTIINKSPEQDGWFAKIELSNESETSQLM TPDAYKKFIQSA*	XP_001299513	*Trichomonas vaginalis*

Table 3.1: List of key enzymes detected in the metabolic pathway of the *P. keilini* hydrogenosome (cont.)

Enzyme Name	Sequence in *P. keilini*	Accession number basted against	Accession number Organism
SHMT Serine hydroxymethyltransferase	>Pseudotrichomonas_keilini_4runs_TRINITY_DN41445_c0_g1_i1.p1_SHMT MEGLELIASENFPSLACLTALSSHFNNKYAEGYPGARYYGGTENVDILERLTQKRALKVF NLNELEWGVNVQALSGSPANFAVYTGLIPPGGKLMGLDLSHGGHLSHGYKTSKKNISSTS LFWKSEPYYVDLNTGLIDYRNLEKKANVFRPNVIIAGTSAYPRHIDYEKMKNISDSTNSI LMSDISHIAGLVAAKIGPNPFLYSDVVTTTTHKTLRGIRGALIFYRKGLNHKTGKEYDYE KKINSAVFPGLQGGPHMHQIAGIAVSLKEALTNDFTQYQKQVVLNMKAMADYLIQNEISL VTGGTDNHLILIDLRPFEIDGVRCQHIFDAVNISTNKNTVPGDSNSFSPKGIRIGSPALT SRGLNENDFRFVGKMIIKGIGITQRINAIAGKNLKKFKELANIDKEIVLLRSIVKRYASS FPLPGIGLHK*	XP_001322593	*Trichomonas vaginalis*
SCS succinyl-coenzyme A synthetase (succinate thiokinase)	>Pseudotrichomonas_keilini_4runs_TRINITY_DN626_c0_g1_i1.p1_SCS MIAQQRFFRNLHKPLLFVDKNTRVVVQGIGNQGQFHSRLMREYGTQVVGAVHPKKAGEII AGLPVFKSVKDCVQKTDANASLIFVPAEGAAQACIDAANSGVGLVVCITEHIPQHDMIRV KKVMKERGVDLIGPNCPGLIQPGTRVKMGIIPTNIHTPGKIGIVSRSGTLTYEAAFATTK AGLGQSTVVGIGGDPFAGQLHTDVIKRFAADPKTEGIILIGEIGGTSEEDAAEWIAKNKL TKEKPIVSFIAGASAPPGKRMGHAGAIVSGGKGTAEGKYQALEAAGIRVARHPGNMGQFI FEEMKRMGRI*	XP_001300482	*Trichomonas vaginalis*
NUOF NADH-quinone oxidoreductase subunit F	>Pseudotrichomonas_keilini_4runs_TRINITY_DN257_c0_g1_i1.p2_NUOF MKRGDWANTDEIVKKGGKWIHGEVKTSEIRGRGGAGFMTGTKWGFLPPVSNKPHYLVINA DEGEPGTCKDRQILTNEPHKLVEGCLLSSMAINAHKCYVYIRGEFSYEAEQLQKAIYEAK DAKLIGKNNKFGWDFDMEIHFGAGAYVCGEETALLNSIEGKAGRPRFKPPFPAIKGLFQC PTIVNNVETISSVPAICKRGGKWFSDIGIPGSKGTKIYGISGHVNHPCVIEDALGVSLKE LIEKHAGGVRGGWDNLLCVIPGGLSCPILSKEEAETAVMGYNELSKMGTALGTGAIIVMD KSTDICKAFDRLSHFYMHESCGQCGPCREGTAWLSEAMSRFAKGKAKRSDLEILEKTSYE TCNCICALAGASSDPIKGLLKHFRKDLEKKLID*	XP_001327980	*Trichomonas vaginalis*
NUOE 24-kDa (NuoE) subunits of the NADH dehydrogenase module in the mitochondrial respiratory complex I	>Pseudotrichomonas_keilini_4runs_TRINITY_DN160340_c0_g1_i1.p1_NUOE GMLSQNGHSSLSLLRFFARVSQKAILIEKEFSFKDQSKIDSIMAKYPSDQPRASIIPLLH LGQRENGGYLTTGVIKAISKITKTPIGRIHETASFYSMFRFSPPNQHIIEICRGLSCYLT GSDNVKTAIEKACDGSFKNKSSKDGLFTLEEVECLGACANAPVMVVNGEYFQNLTAEKAK EIIHRIKAGKSINEFKACNTPPAKPLP*	XP_001312168	*Trichomonas vaginalis*

3.0.2.3 Metabolic pathways in the *P. keilini* hydrogenosome

The way for any substrate into the hydrogenosome is regulated by translocases of the outer and inner membranes. There are several membrane transporters on the hydrogenosomal membranes. On the outer membrane, Tom40 is detected which is part of the Translocase of the Outer Membrane (TOM) complex. The inner membrane contains different transporters such as Tim44, Tim17, Tim16 (Pam 16), Tim14(Pam18), mtHsp70 and Mge1. These substrates are then acted on by a series of enzymes to produce ATP, H_2, and CO_2. Figure 3.1 represents the main metabolic pathways in the hydrogenosome of *P. keilini.*

3.0.2.3.1 Membrane transporters

In canonical mitochondrial organisms, the inner and outer membranes have evolutionarily unrelated translocase complexes, which contain a core translocase and accessory components which assist in preprotein import. The Translocase of the Outer Membrane (TOM) complex has an essential conserved translocase, TOM40, and orthologs have been functionally characterised as the preprotein translocases in isolated hydrogenosomes [122]. Characteristic of these proteins is a beta-barrel fold of the pfam hmm family Porin_3, the number of paralogs of this protein varies from lineage to lineage in excavates with kinetoplastids such a *T. brucei* having two copies of a highly diverged protein termed ATOM[153], and *T. vaginalis* as many as six[154], from our transcriptomic dataset we identify two partial sequences in *P. keilini* (Pfam Porin_3, E-value $1.1x10e-6, 2.3x10e-5$). Whilst neither sequence is complete both have beta barrel topology similar to translocases identified in *T. vaginalis* by PRED_TMBB[13]. In most eukaryotes the TOM complex has accessory proteins which assist in preprotein import and binding, these subunits seem to have independently emerged in different eukaryotic lineages though are assumedly functionally similar. Several accessory proteins have been identified to the *T. brucei* TOM complex[163], though appear absent in the *T. vaginalis* [122]. No strong homologs to the accessory proteins from the yeast system are present in *P. keilini* nor ATOM11, 12, 14, 46, 69 of the *T. brucei* TOM complex.

Preproteins destined for insertion into the outer membrane are handled subsequent to Tom40 import by another bacterially evolved beta barrel protein Sam50 which is the core translocase of the SAM complex[108].

In contrast to the ancestrally bacterial beta barrel translocases of the outer membrane import through and insertion into the inner membrane is facilitated by proteins of eukaryotic innovation and termed Translocase of the Inner Membrane (TIM) complexes. In many eukaryotes the inner membrane translocases have functionally diverged to do slightly different tasks, in yeast three related proteins Tim17, 22, 23 form the core translocases to two different complexes the Tim22 complex mediating

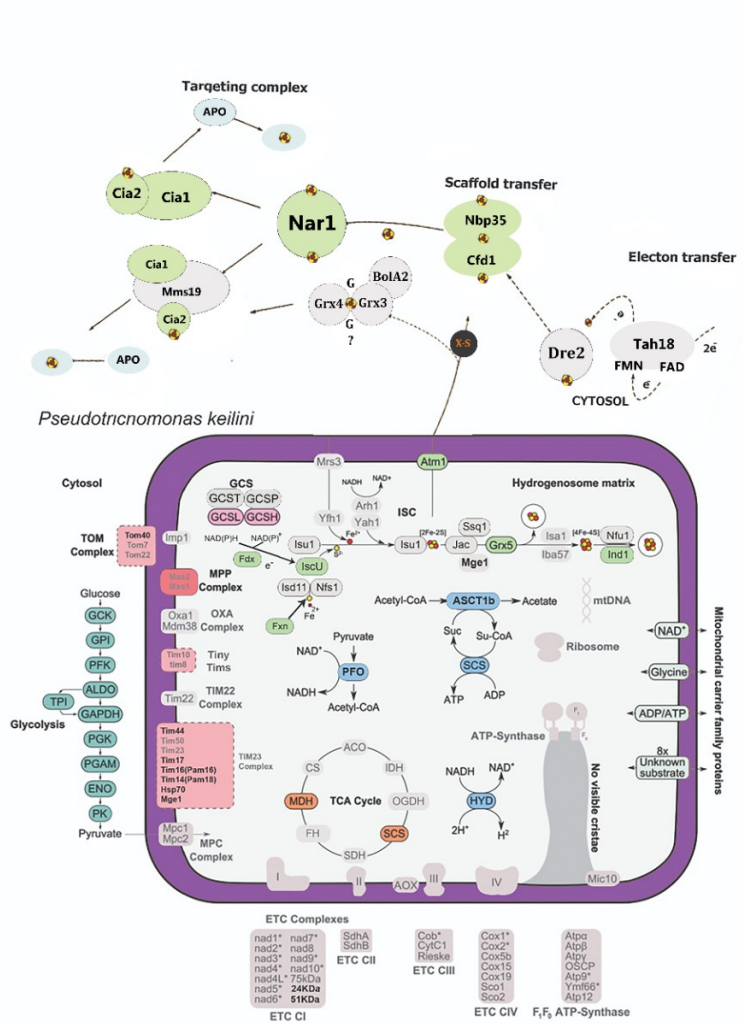

Figure 3.1: Biochemical pathways in the *Pseudotrichomonas keilini* hydrogenosome. This figure illustrates the key reactions essential for energy generation through hydrogeonosomes in *P. keilini* starting with glycolysis which is the main pathway used to produce pyruvate which enters the hydrogenosme through the different membranes transporters TOMs and TIMs, followed by TCA cycle, Glycine Cleavage System (GCS) and the Iron Sulfur Cluster (ISC) pathways which take place both in the cytosol and in the hydrogenosomal matrix. The figure also highlights the ETC subunits that were detected in complex I. Figure design is adapted from Peña-Diaz and Lukeš (2018) and Lewis et al. (2019)[150, 118]

Colored shapes represent that we found a homolog for that enzyme while grey ones indicate missing enzymes.

the insertion of membrane proteins and Tim23 for lumen destined proteins.

We identified two proteins which are homologous to the inner membrane TIM 17 family with E-values of $3.3x10e-8$ and $4.6x10e-8$ in *P. keilini*. This result is comparable to that of *T. vaginalis* where four of the TIM 17 proteins were discovered. However, it is not yet clear whether these proteins form functionally discrete complexes or not[154].

Like the outer membrane, the translocase complexes of the inner membrane have accessory subunits which assist in preprotein import, the most significant of these the Presequence translocase-Associated Motor (PAM) which involves both matrix and membrane associated proteins. The membrane associating components, Tim44, Pam16 (Pfam Pam16, E-value $8.4x10e-5$), and Pam18 were all found in *P. keilini* as well as the matrix proteins Mge1 and mtHSP70. The PAM motor has also been characterised in *T. vaginalis*[154] and is similarly complete.

Some of the imported preproteins use specific Mitochondrial Processing Protease (MPP) to undergo signal sequence cleavage. MPP in many eukaryotes is composed of evolutionarily-related α/β subunits; however, in other eukaryotes such as *T. vaginalis* have undergone genomic streamlining which resuted in a single MPP subunit [198]. In our dataset for *P. keilini*, we found a single type of MPP protease, suggesting that the reductive evolution of this complex occurred before the speciation of *P. keilini*.

3.0.2.3.2 Glycolysis pathway and ATP generation

In order for *P. keilini*'s hydrogenosome to produce energy, the previously mentioned enzymes interact together to form the main metabolic pathways in the hydrogenosome. The pathway starts with the production of pyruvate which is the main substrate that is further used in the catabolic pathway of the hydrogenosome. Pyruvate results from breaking down sugars through the glycolysis pathway. Since carbohydrates were identified as the main source of energy in *T. vaginalis* by Carlton and colleagues in 2007 [38], by analyzing the curated proteins from the dataset, we detected all the key enzymes involved in the glycolysis pathway. As illustrated in Figure 3.1, enzyme list can be found in Table 3.1. Hence, we hypothesize that it is a similar case and that *P. keilini* uses carbohydrates metabolism as a primary pathway for generaing ATPs.

Pyruvate then goes through oxidative decarboxylation by pyruvate:ferredoxin oxidoreductase (PFO) [EC 1.2.7.1] to acetyl-CoA and CO_2 with the accompanying reduction of ferredoxin whose electrons are transported to protons generating H_2 via the [FeFe]-hydrogenase (hydA) activity.

In order for a mature [FeFe]-hydrogenase (hydA) to be assembled correctly , three maturase enzymes need to be involved. These enzymes are: small GTP-binding protein (hydF), FeFe-hydrogenase assembly protein (hydG) and radical SAM domain containing protein (hydE).

The resulting acetyl-CoA is then acted on by ASCT1B which catalyzes the bidirectional reaction of acetyl-CoA and acetate.

3.0.2.3.3 Krebs cycle

Hydrogenosomes are generally known to lack a complete TCA cycle [52, 92, 21], especially in the best-studied ones of *T. vaginalis*. While the Krebs cycle is incomplete in the hydrogenosome of *P. keilini*, the malate dehydrogenase enzyme seems to have been retained despite the reductive evolution. This enzyme acts as a catalyst in the reversible oxidation of malate to oxaloacetate and uses the reduction of NAD^+ to NADH. Since malate dehydrogenase is involved in other metabolic pathways, the detection of it on the bioinformatic level does not necessarily mean that it is active in the Krebs cycle pathway. Another Krebs cycle enzyme which was also detected, is Succinyl coenzyme A synthetase (SCS) which catalyzes the inter-conversion of succinyl-CoA to succinate.

3.0.2.3.4 Components of the electron transport chain (ETC)

Since hydrogensosomes do not produce ATP by oxidative phosphorylation (OXPHOS) complexes, they lack the key enzymes and subunits that are necessary for that process. However, some of these parts seem to have been retained in the hydrogenosome of *P. keilini* despite the reduction process that has occurred during their evolution. These components include 51-kDa NADH-quinone oxidoreductase subunit F (NUOF) and 24-kDa (NUOE) subunits of the NADH dehydrogenase module in the mitochondrial respiratory complex I. Both enzymes are also present in the hydrogensosome of *T. vaginalis*. One of the reasons why the discovery of these subunits was surprising, was due to the fact that hydrogenosomes lack cytochromes [138]. Their discovery has solved the missing chain in how NAD^+ are regenerated after malate oxidation. In their paper, Hardy and colleagues (2014) showed through their phylogenetic analyses that both hydrogenosomes and mitochondria are merely aerobic and anerobic forms of the same organelle which was endosymbiotically acquired.

Our result conforms with that of the Hadry and colleagues (2014). We found that the proteins sequences of *P. keilini*'s hydrogenosomal enzymes also do not branch within these of the alphaproteobacteria with the exception of GCSH and NUOF trees. It was also interesting to find that the proteins of *T. vaginalis* had longer branches than the ones of *P. keilini*.

Mitochondrial complex I is the fist enzyme that acts in the respiratory chain. It starts by oxidizing NADH, a product of the krebs cycle which takes place in the mitochondrial matrix. The two electrons that result from that process are then used for the reduction of ubiquinone to ubiquinol. Therefore, complex I is considered the entry point by which electrons enter the respiratory chain and are then transferred between different complexes that then leads to the conversion of ADP to ATP.

The 51-kDa and 24-kDa subunits are parts of the NADH dehydrogenase module in complex I. It was discovered that these enzymes in the hydrogenosome of *T. vaginalis* have similar propereties to the ones normally found in mitochondria in the reduction of electron carriers; however, they were found to be also capable of ferredoxin reduction, the electron carrier which is used for hydrogen production [92].

3.0.2.3.5 Iron-sulfur cluster (ISC) pathway

The Fe-S pathway is considered the only conserved one across all mitochondrial and mitochondrial-related organelles. It is also considered the main pathway which is vital for the survival of organisms in the mitochondria. It has been revealed that mutant yeast lacking the oxidative phosphorylation could still survive and keep their mitochondria if there's a carbon source available for growth[137]; however, they were unable to do so when they lacked the genes coding for the ISC [120].

Using comparative analysis of the ISC enzymes from *T. vaginalis* and *Naegleria gruberi*, we detected most of the ISC enzymes of *T. vaginalis* in the transcriptome of *P. keilini*.

There are two main pathways necessary for the formation of the FeS cluster in eukaryotes. The first is the ISC which is found in the mitochondria or mitochondrial-related organelles (hydrogenosomes in the case of *P. keilini*), and the second is the Cytosolic Iron-sulfur protein Assembly (CIA) pathway which takes place in the cytosol.

Until recently, *T. vaginalis* was thought to be the only excavate that contains a homolog of the iron–sulfur flavoprotein (Isf) which is an enzyme of a bacterial-type [150]. This enzyme is involved in the detoxification of Reactive oxygen species (ROS) and in protecting the organism from the rise of oxygen levels in the environment. However, we have detected a homolog of the Isf enzyme in the transcriptome of *P. keilini* as well. Given that both *T. vaginalis* and *P. keilini* are microaerophiles that live in low or oxygen-free environments, it seems that harbouring this enzyme is a survival necessity for such organisms.

In addition to the Isf enzyme, we also detected most the other ISC and CIA enzymes which are present in *T. vaginalis*. The ISC enzymes are IscU, IscS, and Fxns,[2Fe–2S] Fdxs, Isa2, Nfus, Grx5, and Ind1s as well as chaperone Hsc20. As for the CIA ones, they are; Cfd1, Nbp35, Cia1 and Cia2 with the exception of Isd11

being missing from *P. keilini*.

3.0.2.3.6 Glycine cleavage system

A half-complete glycine-cleavage system was also detected in *P. keilini*. That system originally includes four components which associate loosely. These components are T-protein (GCST), P-protein (GCSP), L-protein (GCSL), and H-protein (GCSH), which do not form a stable complex, they are rather referred to as a system that catalyzes a series of reversible reactions. The direction of the reaction is determined in response to the amino acid glycine concentration.

A complete reaction for the gylycine-cleaveage system (GCS) goes as follows: $\mathrm{Glycine} + \mathrm{H_4folate} + \mathrm{NAD^+} \rightleftharpoons 5{,}10-\mathrm{methylene}-\mathrm{H_4folate} + \mathrm{CO_2} + \mathrm{NH_3} + \mathrm{NADH} + \mathrm{H^+}$

The mechanism of action for this system starts first by H-protein which activates the P-protein, which then catalyzes glycine decarboxylation and then attaches the intermediate molecule to the H-protein. That molecule will then be shuttled to the T-protein [70, 146]. The H-protein and the T-protein form a complex that uses tetrahydrofolate and results in ammonia and 5,10-methylenetetrahydrofolate.

Following the interaction with the T-protein, the resulted products are H-protein and two fully reduced thiol groups as part of the lipoate group [69].

As for the regeneration of the glycine protein, it is done by the oxidation of the H-protein which interacts with the L-protein to regenerate the disulfide bond in the active site and reduces NAD+ to NADH and H^+.

However, the parabasalid under study *P keilini* lacks a complete glycine-cleavage system, which is also the case for its close-relative *T. vaginalis* [172].

The first component that is missing in both of them is the P protein (EC 1.4.4.2). That enzyme is responsible for converting glycine and [glycine-cleavage complex H protein]-N6-lipoyl-L-lysine to [glycine-cleavage complex H protein]-S-aminomethyl-N6-dihydrolipoyl-L-lysine with the release of CO_2 in the process [83, 148, 140].

As for the second component which is the T protein (EC 2.1.2.10, aminomethyl-transferase), known as glycine synthase, also seems to be missing from *P. keilini*. The main function of it is catalyzing the conversion of [protein]-S8-aminomethyldihydrolipoyllysine and tetrahydrofolate to [protein]-dihydrolipoyllysine and 5,10-methylenetetrahydrofolate and NH_3

On the other hand, two of the main proteins were identified in *P. keilini* such as the L protein (EC 1.8.1.4, dihydrolipoyl dehydrogenase) and the lipoyl-bearing H protein.

Another enzyme that was also present is the Serine hydroxymethyltransferase (SHMT) (EC 2.1.2.1) ,SHMT catalyzes the interconversion of glycine to serine as follows:

5,10−methylenetetrahydrofolate + glycine + H_2O = tetrahydrofolate + L−serine [6, 27, 68, 111, 162].

To conclude the findings on the metabolism of the *P. keilini* hydrogenosome, we can say that it retains the core translocases of the inner and outer membranes, and resembles closely the architecture of the Trichomonas hydrogenosome, with complete PAM and a single MPP, but without other accessory proteins. The copy number of the core translocases is lower than that of *T. vaginalis*, particularly interesting are the two Tim17 family proteins, which would suggest that the scope for functional diversity of the TIM complexes is limited (to those two proteins) and could be another example of a reductive evolution in the preprotein import system. It also retains the main hydrogenosomal enzymes acting in the energy production pathway in *T. vaginalis*. The hydrogenosome of *P. keilini* seems to be highly similiar to the one in *T. vaginalis*, but not identical as as it is missing the ASCT subunit C and retained the ASCT subunit B which is missing in *T. vaginalis*' hydrogenosome.

3.0.2.4 Phylogenies of the hallmark enzymes

To understand the evolution of hydrogenosomes in Parabasalids we constructed phylogenetic trees for hallmark enzymes which included sequences from Stairs et al.(2015) from all organisms containing any of the forms of mitochondria and its related orgnelles. Using mafft v7.390 [104] and BMGE-1.12 [44] for multiple sequence alignments, we constructed trees using IQ-TREE-v1.6.10 [141] LG+C60 model.

By analyzing the tree topology of the different hydrogenosomal proteins, we could see that most key enzymes that are present in the hydrogenosome of *T. vaginalis* according to Stairs et al.(2015) are also detected in *P. keilini*. Following the construction of phylogenetic trees for the 13 key enzymes identified in *P. keilini* 2.0.6.1, in all of the trees where homologs were found for the three parabasalids *T. vaginalis , T. foetus* and *P. keilini*, the three of them were grouping together. Given the different lifestyles of all of them, ranging from parasitic to free-living styles, it was rather interesting to find them together in one clade

> given the differences in their lifestyles we expected *P. keilini* to perhaps group with other free-living eukaryotes instead as it would indicate a common ancestor for free-living eukaryotes rather than a transition in lifestyle for parabasalids.

Hence, the trees topology support the hypothesis that the common ancestor of parabasalids had a hydrogenosome that was passed on to these organisms (see figures 3.2 to 3.14).

Figure 3.2: Phylogeny of Pyruvate:ferredoxin oxidoreductase (PFO), which is a key hydrogenosomal enzyme that catalyzes the interconversion of pyruvate to Acetyl-CoA. Most sequences of excavates *Psalteriomonas lanterna,* and *Sawyeria marylandensis* are grouping together with the parabasalids *P. keilini, T. vaginalis,* and *T. foetus*

3.0.3 Comparative Genomics

In order to better understand the evolutionary processes that *P. keilini* has undergone, we used Orthofinder to find the closest homologs between *P. keilini* and *T. vaginalis*. We found 12,202 proteins of *P. keilini* corresponding to 13,504 orthologs in *T. vaginalis*. That means that 9,098 proteins in *P. keilini* have no homologs in *T. vaginalis* which has 37,265 unique proteins that do not match any in *P. keilni*. That could be due to the genome duplication of *T. vaginalis*, or because of the over-conservative method that was used in filtering the transcriptomic data in *P. keilini* which may have lead to missing some of its *bona fide* proteins. Another explanation is that it was caused by a lineage specific evolutionary event such as Horizontal Gene Transfer (HGT).

3.0.4 Metabolic pathways in *P. keilini*

In this section, we analyzed the transcriptomic dataset for *P. keilini* to determine the main energy source that a free-living parabasalid such as *P. keilini* uses. As with the previous analysis, we compared the transcriptomic data with that for close relatives like *Trichomonas vaginalis* to try and determine the metabolic capabilities for *P. keilini*.

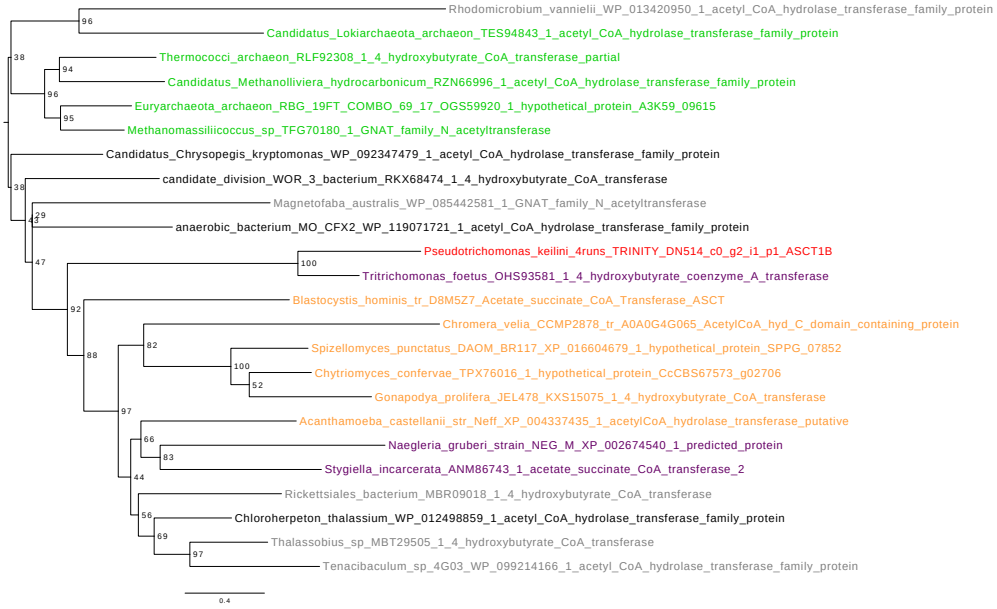

Figure 3.3: Phylogenetic analysis of Acetyl:succinate CoA-transferase subunit b (ASCT1b) which is a bidirectional enzyme that acts on the conversion of succinate to acetyl-CoA and vice versa. In the tree, *P. keilini* is grouping with *T. foetus* and they are both forming a bigger clade which includes eukaryotic sequences, some of which are excavates.

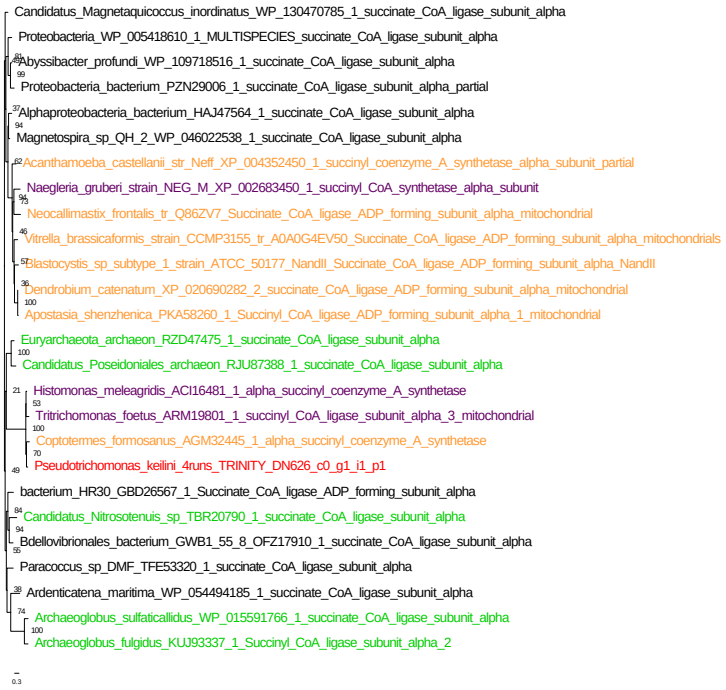

Figure 3.4: Phylogenetic analysis of Succinyl coenzyme A synthetase (SCS) which is a Krebs cycle enzyme that catalyzes the interconversion of succinyl-CoA to succinate. We can see that eukaryotic sequences of the enzyme are not grouping together; however, the parabaslid sequences such as those of *T. foetus, T. vaginalis*, and *P. keilini* are forming a clade together.

Figure 3.5: Phylogenetic analysis of Malate dehydrogenase, the second Krebs cycle enzyme that we detected in *P. keilini*. It stimulates the oxidation process by which malate is converted to oxaloacetate, which leads to the production of NADH from NAD^+ by reduction. In the tree, parabasalid sequences of *P. keilini, T. vaginalis, T. foetus,* and *T. gallinae* are grouping in a clade, while sequences of other excavates like *Naegleria gruberi, Spironucleus salmonicida,* and *Giardia intestinalis* are not grouping together.

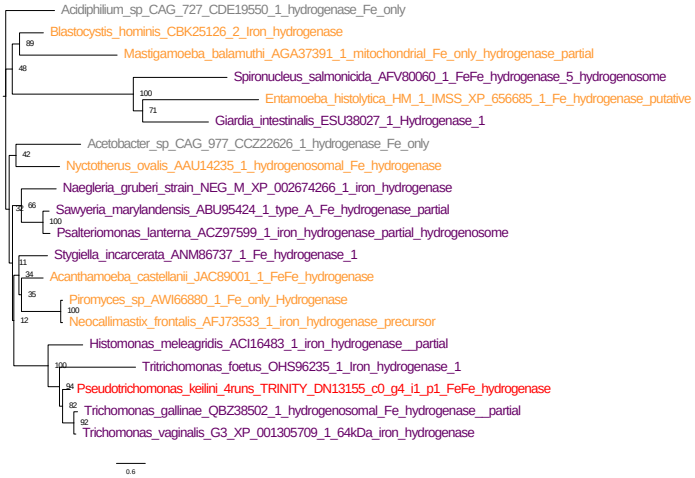

Figure 3.6: Phylogeny of Fe-Fe hydrogenase (hydA) enzyme which is responsible for the production of molecular hydrogen and one of the key hydrogenosomal enzmes. Another sequence from the *Histomonas meleagridi* excavate is grouping with the parabasalian clade of *P. keilini, T. vaginalis, T. foetus, Trichomonas sp* and *T. gallinae*.

Figure 3.7: Phylogeny of radical SAM domain containing protein (hydE), one of the three maturase enzymes required for the synthesis of a mature Fe-Fe-hydrogenase(hydA). In this tree, the three parabasalids of *P. keilini, T. vaginalis,*and *T. foetus* are grouping together.

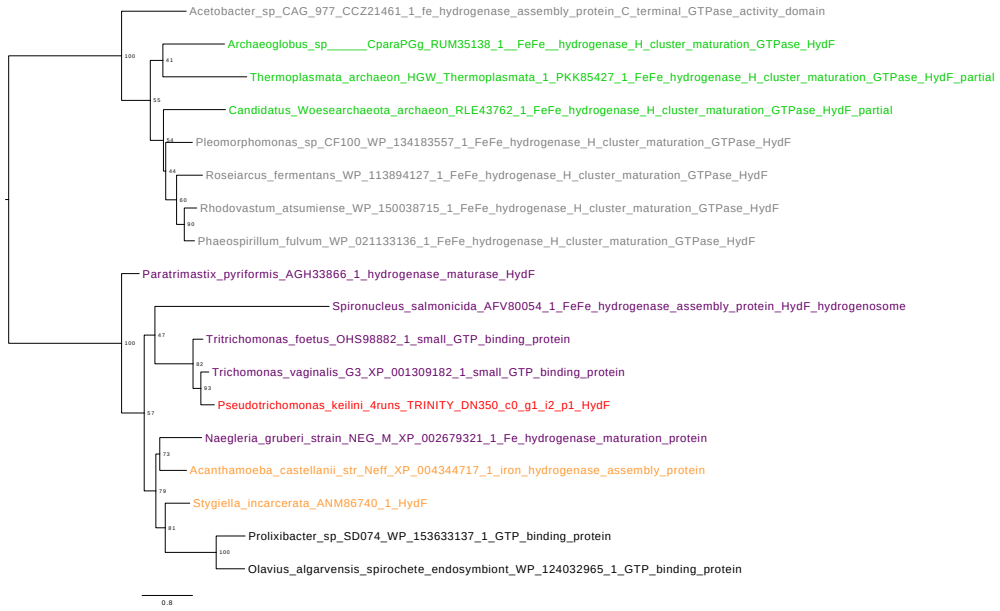

Figure 3.8: Phylogeny of small GTP-binding protein (hydF), another one of the three maturase enzymes required for the synthesis of a mature Fe-Fe-hydrogenase(hydA). In this tree, the three parabasalids of *P. keilini, T. vaginalis,* and *T. foetus* are grouping together and eukaryotes seem to be grouping together except for *Stygiella incarcerata* which is grouping with bacterial sequences.

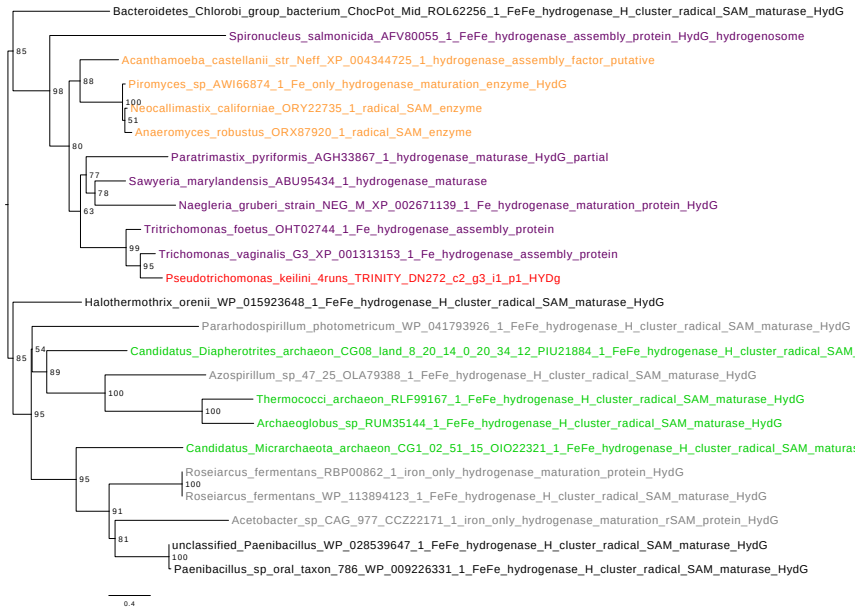

Figure 3.9: Phylogeny of FeFe-hydrogenase assembly protein (hydG), the third maturase enzyme required for the synthesis of a mature Fe-Fe-hydrogenase(hydA). A clade of the three parabasalids *P. keilini, T. vaginalis,* and *T. foetus* can be seen which is part of a bigger clade of eukaryores including excavates. It is worth noting that the bacterial sequence from *Bacteroidetes Chlorobi group bacterium* is branching with this group.

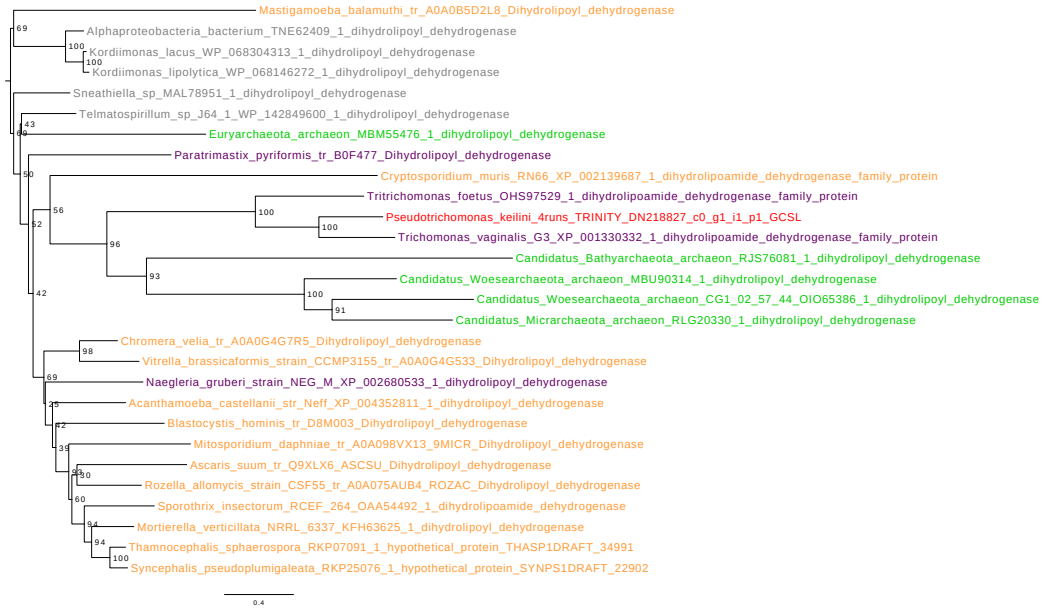

Figure 3.10: Phylogeny of L-protein (GCSL) which is part of the glycine cleavage system. The tree topology shows a clade of parabaslids consisting of *P. keilini, T. vaginalis,* and *T. foetus*, while other eukaryotic sequences form a separate clade.

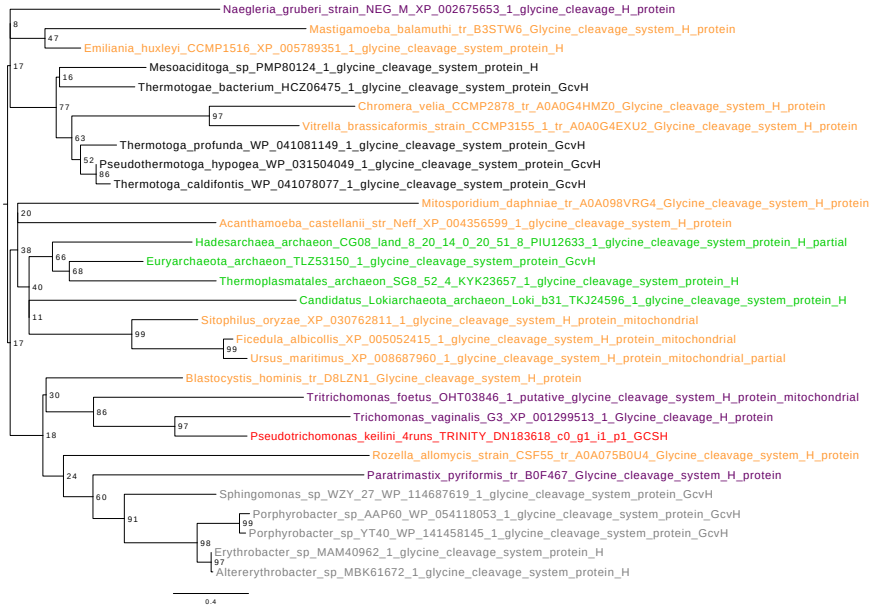

Figure 3.11: Phylogeny of H-protein (GCSH), the second enzyme detected in *P. keilini*'s glycine cleavage system. A clade of parabaslids involving *P. keilini, T. vaginalis,* and *T. foetus* can be seen which is part of a larger clade composing other eukaryotic sequencs and some alphaproteobacterial ones which could indicate an alphaproteobacterial origin of the enzyme

Figure 3.12: Phylogenetic analysis of Serine hydroxymethyltransferase (SHMT) enzyme which also plays a role in the glycine system. The tree shows that neither eukaryotes nor excavates are monophyletic. However, *P. keilini*and *T. vaginalis* are grouping together.

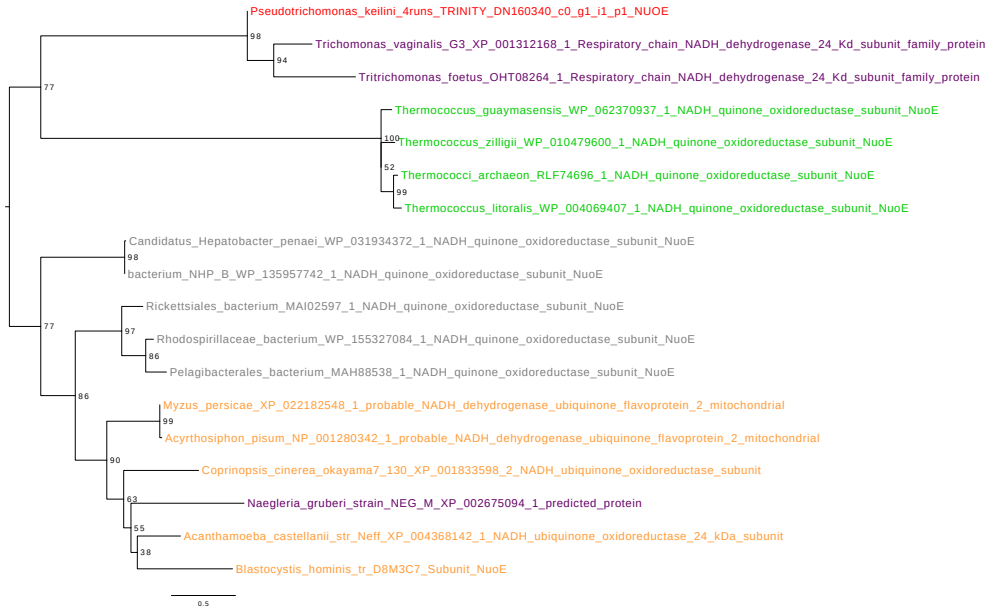

Figure 3.13: Phylogeny of 24-kDa NADH-quinone oxidoreductase subunit E (NUOE) which is one of the components of Complex I of the electron transport chain. In this tree, *P. keilini* sequence is grouping with the other parabasalian ones of *T. vaginalis* and *T. foetus*, and not with alphaproteobacteria.

Figure 3.14: Phylogenetic tree of 51-kDa NADH-quinone oxidoreductase subunit F (NUOF) , another component of Complex I of the electron transport chain. The tree topology shows that the parabasalian group of *P. keilini, T. vaginalis,* and *T. foetus* are grouping together within a larger clade of alphaproteobacterial sequences. This could indicate an endosymbiotic origin of the enzyme.

3.0.4.1 Lipid metabolism

Being a bacteriovore, we hypothesized that *P. keilini* would have the enzymatic capabilities to break and metabolize the phospholipd bilayer of bacteria and harness it for energy production. The starting point of research was to search the identified set of lipid-metabolizing enzymes of *Naegleria gruberi* against the proteome of *P. keilini*. *N. gruberi* is a free-living excavate that prefers lipids as a substrate for energy production [20]. The phylogenetic relationship between both *P. keilini*, and *N. gruberi* gave more support to this hypothesis.

The main question was whether *P. keilini* prefers lipids as a substrate like *N. gruberi* since they are both free-living excavates, or if carbohydrates are the main preferred energy source such as its parasitic relative *T. vaginalis*. The BLAST search for the identified lipid metabolism enzymes in *N. gruberi* against the proteome of *P. keilini* resulted in only 24% approximately in *P. keilini* of the queried enzymes. This indicates that the metabolism of *P. keilini* is more similar to its parasitic relatives than that of its free-living ones.

3.0.4.2 Amino acid metabolism

Several amino acids were identified to be used as energy substrate in *T. vaginalis* [38]. However, genes coding for key enzymes in the serine synthesis pathway were missing in *T. vaginalis* [38]. These enzymes are phosphoserine phosphatase (EC: 3.1.3.3) and serine-pyruvate transaminase (EC:2.6.1.51), which synthesises serine from hydroxypyruvate. Through analyzing the annotated transcriptomic data, we identified phosphoserine phosphatase (EC: 3.1.3.3) in *P. keilini*.

Furthermore, another pathway which is missing in *T. vaginalis* is the Methionine regeneration pathway which is responsible for the conversion of Methylthioadenosine to Methionine [38]. However, we identified one of the enzymes in that pathway in *P. keilini* which is 5'-methyladenosine nucleosidase EC 3.2.2.9. That enzyme acts on S-methyl-5'-thioadenosine to produce adenine and S-methyl-5-thioribose [66].

3.0.5 Phylogenetic position of *P. keilini*

To investigate the relationship of *P. keilini* to other parabasalids and metamonads, we inferred a maximum likelihood combined protein phylogeny under the best-fitting LG+C60+F+R6 model from a dataset of 57 single-copy orthologous genes conserved on at least 75% of a representative sample of 21 parabasalid, metamonad and discoban genome or largely-complete transcriptome datasets, with 9 additional eukaryotes included as an outgroup (3.15). Interestingly, the phylogeny indicates that *P. keilini* branches within a clade of parasitic parabasalids with maximal bootstrap support, as the sister lineage to a clade comprising *Trichomonas vaginalis*, *T.*

gallinae and *T. tenax*. The cattle and cat parasite *Tritrichomonas foetus* forms an outgroup to this *P. keilini*-*T. vaginalis* clade (3.15). The phylogeny implies either than *P. keilini* evolved a free-living lifestyle from a parasitic ancestor, or alternatively that there have at least two transitions to parasitism within this clade — once in the ancestor of *T. foetus* and once in the ancestor of the *T. vaginalis* clade.

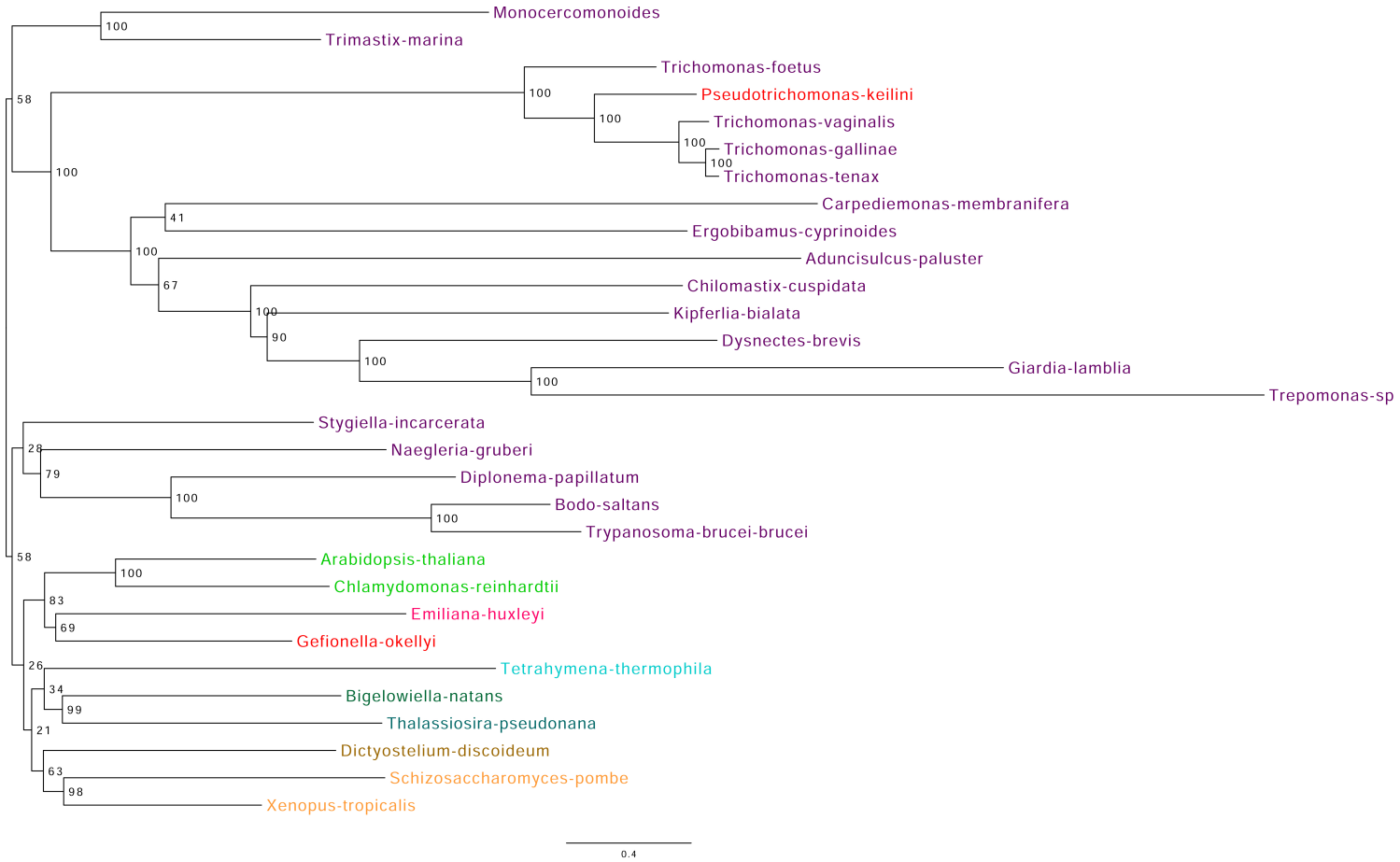

Figure 3.15: Unrooted Species tree of *Psuedotrichomonas keilini* among parabasalids and metamonads. We inferred a maximum likelihood combined protein phylogeny under the best-fitting LG+C60+F+R6 model from a dataset of 57 single-copy orthologous genes conserved on at least 75% of a representative sample of 21 parabasalid, metamonad and discoban genome or largely-complete transcriptome datasets, with 9 additional eukaryotes included as an outgroup We can see very high bootstrap support values for *P. keilini* grouping with the other parasitic parabasalids *T. vaginalis, T. foetus, T. gallinae*, and *T. tenax.*

3.0.5.1 Phylogenetic trees of conserved eukaryotic proteins

Williams and colleagues (2017) identified 44 conserved proteins across eukaryotes [189]. We used the HMM alignment profiles for these proteins to do an HMM search against the proteome of *P. keilini* and we identified 41 proteins that matched the threshold e-value score of the alignment(0.00001). In addition, we constructed phylogenetic trees for these proteins following the method in 2.0.6.1 using mafft and BMGE for alignment [104][44]. We then used I Q-TREE-v1.6.10 [142](Supplementary Figures 3.17 to 3.57).

The only three missing markers are i)Tetrapyrrole (Corrin/Porphyrin) Methylase (Wlm17007_OG11907.aln), ii) Protein kinase superfamily protein (Wlm17018_OG18944.aln) and iii) RNAse l inhibitor protein 2 (Wlm17043_OG842.aln). On the other hand, the identification of this large number of markers is an indicator for the completeness of the *P. keilini* transcriptome.

3.0.5.1.1 Phylogenetic position of *P. keilini* using supermatrix analysis

Using the phylogenetic trees for the eukaryotic marker proteins, we conducted a supermatrix analysis for the purpose of producing another species tree that can then be compared with the one we constructed using gene orthologs 3.15. For this analysis, we used a homemade script to unify the leaves names in all of the 41 alignments files to only show the species names. Following that, we applied the alignment concatenation method in IQ-TREE -v 2.0.5 [141] which resulted in the following species tree.

Figure 3.16: Species tree inferred from the concatenated alignment of the 41 eukaryotic marker proteins also using the LG+C60+F+R6 model. The tree confirms the position of *P. keilini* among the other parasitic parabasalids *T. vaginalis*, and *T. foetus*. However, what is interesting in this tree topology is the position of *Monocercomonoides*. According to the species tree 3.15, it is evolutionarily closer to the metamonad *Trimastix marina* than to the other parabasalids, while its position in this tree of eukaryotic markers, it appears to be closer to the parabasalian clade than to *T. marina*.

Figure 3.17: Phylogenetic tree of the ribosomal protein Rp L16p/ L10e (Wlm17001 alignment). The three parabasalids *P. keilni, T. vaginalis,* and *T. foetus* are forming a clade with a long branch on the tree. *Monocercomonoides* forms an outgroup to that clade.

Figure 3.18: Phylogenetic analysis of the DNA ligase enzyme (Wlm17002 alignment). *P. keilini* sequence is grouping with that of *T. vaginalis* and *T. foetus* with maximum support value. The tree topology also shows that the three parabasalids are branching with other excavates like *Trepomonas sp, Dysnectes brevis,* and *Monocercomonoides*. Other excavates are grouping with other eukaryotic sequences elsewhere in the tree.

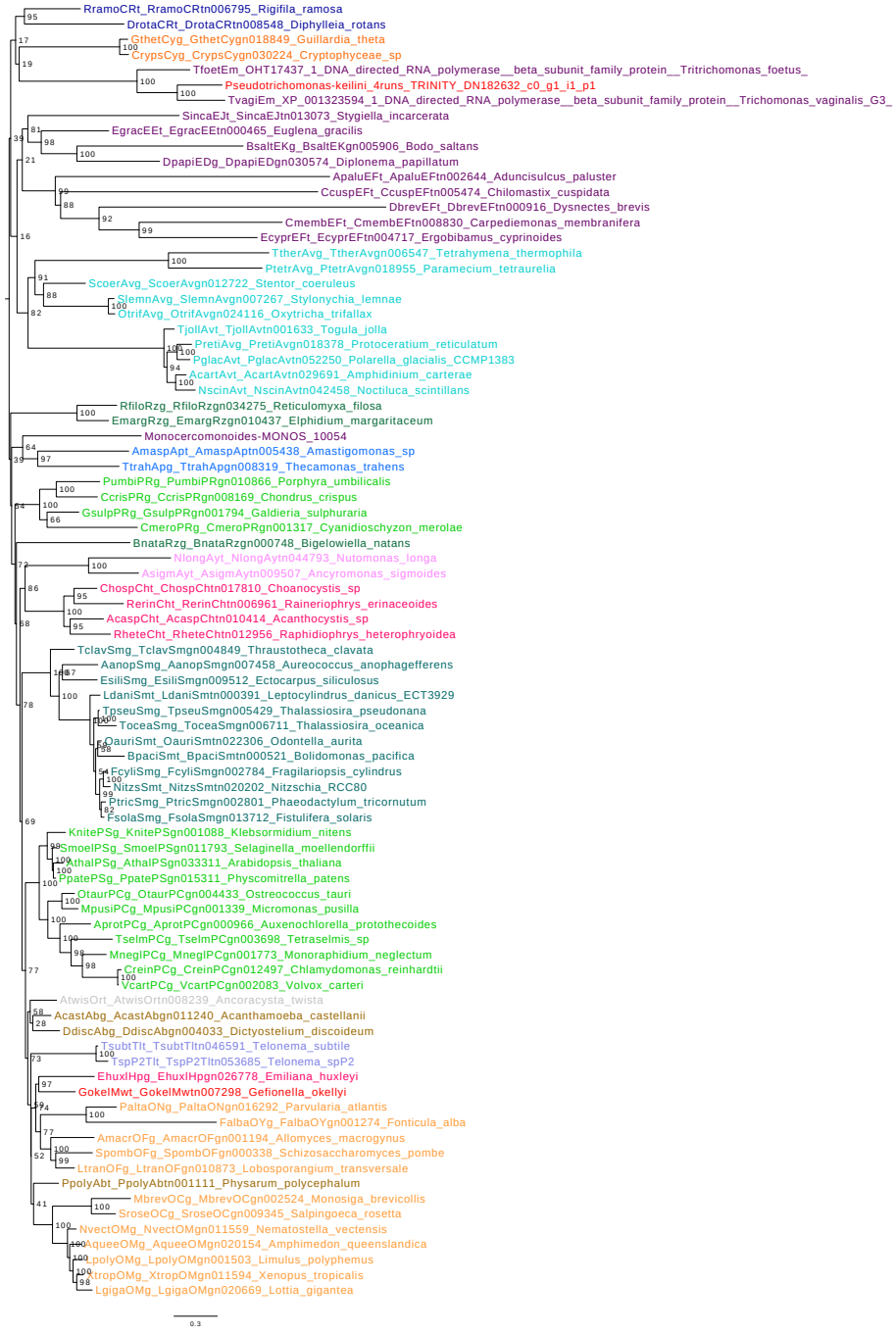

Figure 3.19: Phylogenetic analysis of the RNA polymerase II subunit 2 (Wlm17003 alignment). In this tree, the three parabasalids; *P. keilini*, *T. vaginalis* are forming a clade with maximum support value and *T. foetus* is an outgroup to that clade.

Figure 3.20: Phylogeny of Elongator factor complex protein 3 (Wlm17004 alignment). The tree topology interestingly shows monophyly of excavates that includes *P. keilini, Stygiella incarcerata, Monocercomonoides, T. foetus, T. vaginalis, Carpediemonas membranifera, Aduncisulcus paluster, Chilomastix cuspidata, Trepomonas sp,* and *Dysnectes brevis*

Figure 3.21: Phylogeny of ATPase, V1 complex, subunit B protein (Wlm17005 alignment). In this tree, parabaalids are not branching with other excavates , while *P. keilini,* and *T. vaginalis* form a clade and *T. foetus* is an outgroup.

Figure 3.22: Phylogeny of 40S ribosomal protein S9-1 (Wlm17006 alignment). In this tree, parabaalids are not branching with other excavates , while *P. keilini*, and *T. vaginalis* form a clade and *T. foetus* is an outgroup.

Figure 3.23: Phlogeny of Ribosomal protein L23/L15e family protein (Wlm17008 alignment). In this tree, most excavates appear to be grouping together including the three parabasalids *P. keilini, T. vaginalis,* and *T. foetus* are forming a clade together.

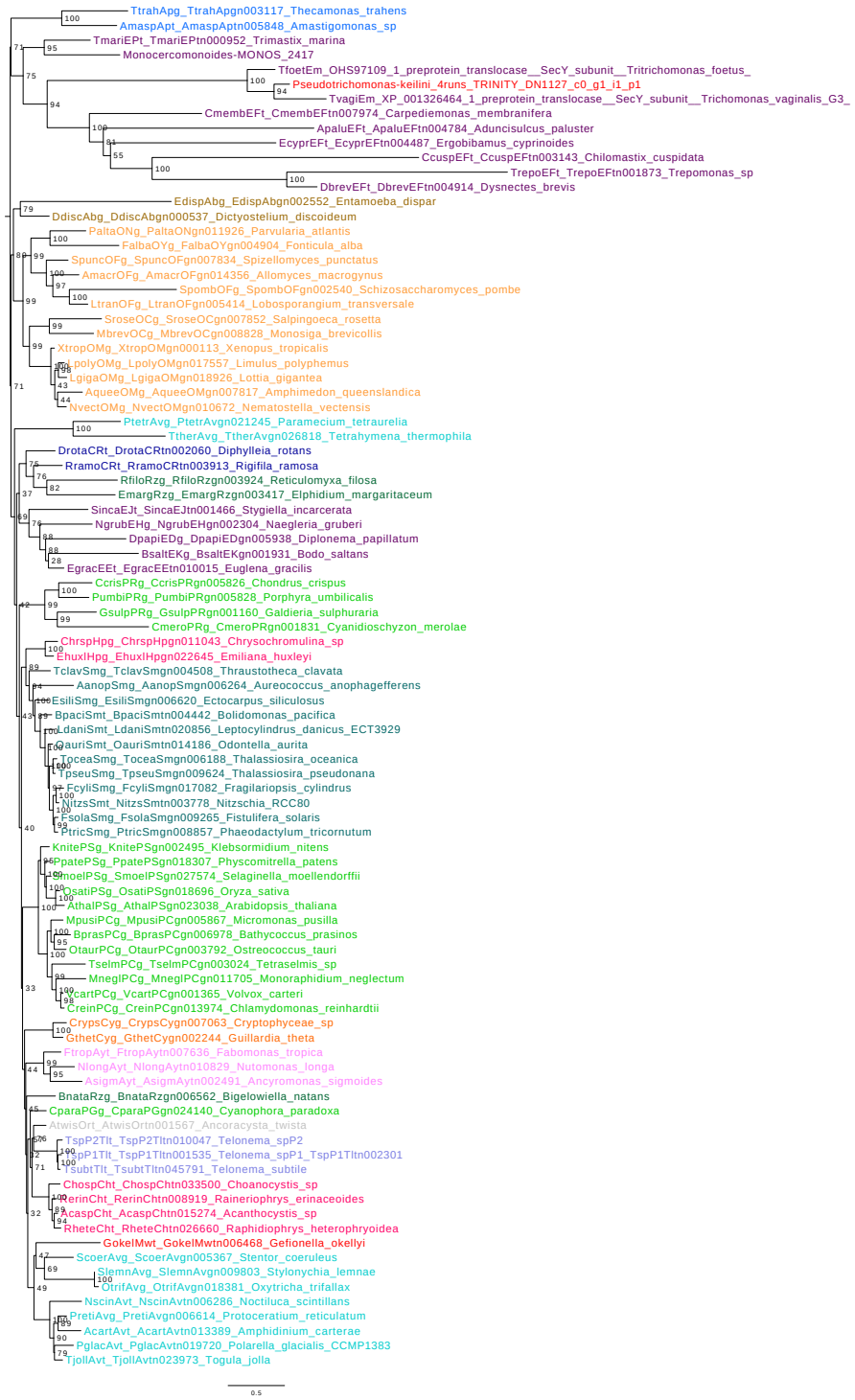

Figure 3.24: Phlogeny of SecY protein transport family protein (Wlm17009 alignment). In this tree, most excavates appear to be grouping together including the three parabasalids *P. keilini, T. vaginalis,* and *T. foetus* are forming a clade together.

Figure 3.25: Phylogeny of Signal recognition particle, SRP54 subunit protein in eukaryotes (Wlm17010 alignment). In this tree, parabasalids are not grouping with the other excavates in the tree. *P. keilini, T. foetus,* are forming a clade and *T. vaginalis* is the outgroup.

Figure 3.26: Phylogeny of ribosomal protein large subunit 16A (Wlm17011 alignment). In this tree, parabasalids form the longest branch in the tree, and interestingly, *P. keilini* is an outgroup to the Trichomonas clade of *T. vaginalis*, and *T. foetus*.

Figure 3.27: Phylogeny of Ribosomal protein L22p/L17e family protein (Wlm17012 alignment). The excavate *Bodo saltans* is an outgroup to the parabasalian clade of *P. keilini*, *T. foeus*, and *T. vaginalis*.

Figure 3.28: Phylogeny of eukaryotic translation initiation factor 2 gamma subunit (Wl17013 alignment). Parabasalids are branching with other eukaryotes. In the clade of *P. keilini,* and *T. vaginalis*, *T. foetus* is the outgroup.

Figure 3.29: Phylogeny of elongation factor 1-alpha in eukaryotes (Wm17014 alignment). The excavate *Naegleria gruberi* forms an outgroup to the parabasalids *P. keilini, T. foetus,* and *T. vaginalis.* By taking a closer look at the parabasalian clade, we can see that the sequence from *T. vaginalis* is the outgroup to the other parabasalids.

Figure 3.30: Phylogeny of eukaryotic release factor 1-2 (Wlm17015 alignment). The three parabasalids *P. keilini, T. foetus,* and *T. vaginalis* are grouping with six other excavates. In the parabasalian clade, *T. foetus* represents the outgroup.

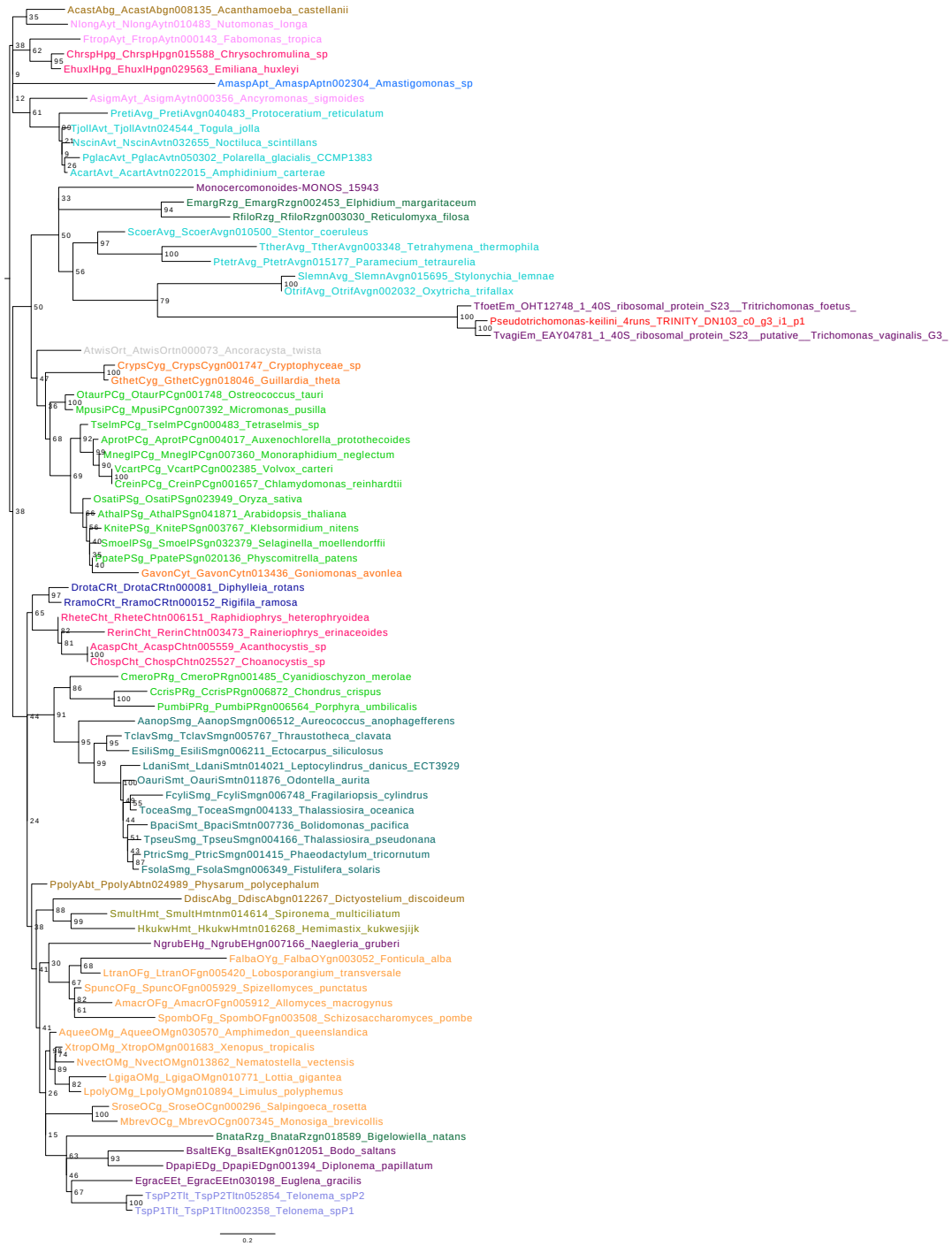

Figure 3.31: Phylogeny of Ribosomal protein S12/S23 family protein (Wlm17016 alignment) where parabasalids form the longest branch on the tree. In their clade, *T. foetus* is the outgroup to *P. keilini,* and *T. vaginalis.*

Figure 3.32: Phylogeny of Ribosomal protein S13/S18 family (Wlm17017 alignment). Parabasalids form the longest branch on the tree, and they group with three other excavates. *Trepomonas sp* represents the outgroup of the parabasalian clade of *T. vaginalis, T. foetus,* and *P. keilini.*

Figure 3.33: Phylogeny of tRNA synthetase beta subunit family protein (Wlm17019 alignment). Parabasalids group with two other excavates *Dysnectes brevis,* and *Trepomonas sp* which form an outgroup to the parabasalian clade of textitT. vaginalis, T. foetus, and *P. keilini.*

Figure 3.34: Phylogeny of cytosolic ribosomal protein S15 in eukaryotes (Wlm17020 alignment). The parabasalian clade of *P. keilini,* and *T. vaginlis* form the longest branch on the tree with *T. foetus* being the outgroup.

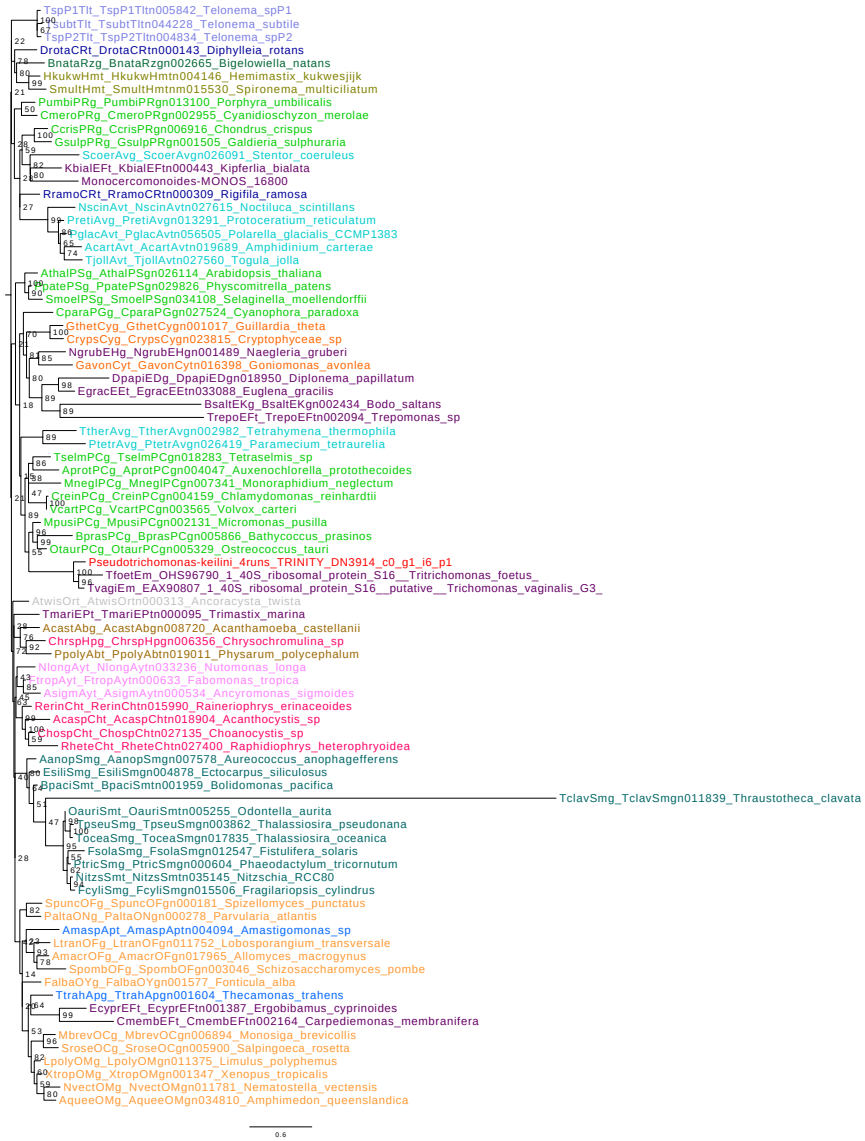

Figure 3.35: Phylogeny of Ribosomal protein S5 domain 2-like superfamily protein (Wlm17021 alignment). *P. keilini* is an outgroup to the clade of Trichomonas *T. vaginalis*, and *T. foetus*. The three parabasalids are not grouping with the excavates in the tree.

Figure 3.36: Phylogeny of Translation protein SH3-like family protein (Wlm17022 alignment). Parabasalids form the longest branch on the tree. *Monocercomonoides* is an outgroup to the parabasalian clade while *T. foetus* is an outgroup to *P. keilini,* and *T. vaginalis.*

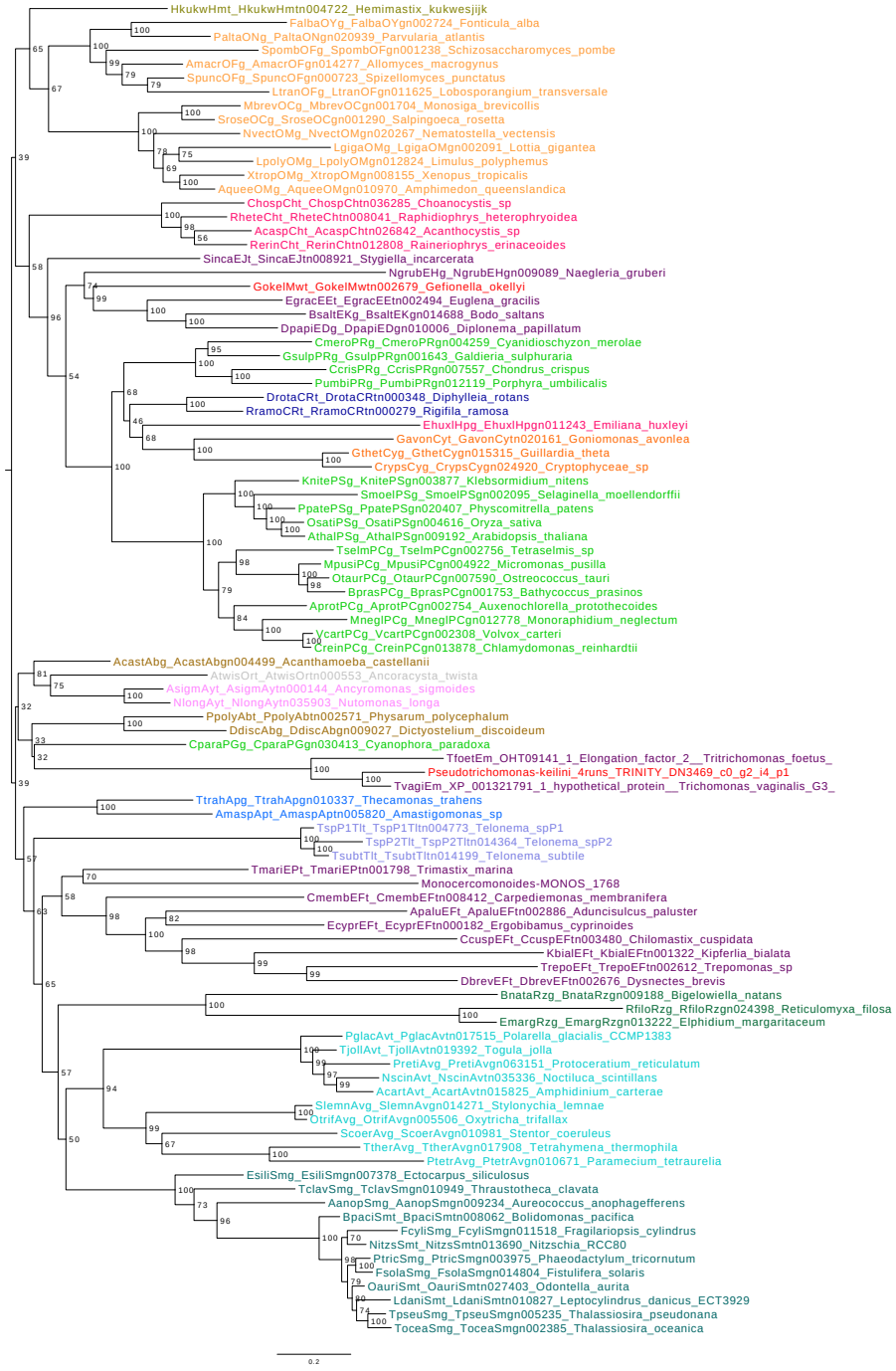

Figure 3.37: Phylogeny of elongation factor EF-2 in eukaryotes (Wlm17023 alignment). The parabasalids *P. keilini*, and *T. vaginalis* are grouping together with *T. foetus* being the outgroup. The excavates in the tree are not monophyletic.

Figure 3.38: Phylogeny of Ribosomal protein S19e family protein (Wlm17024 alignment).
The excavate *Monocercomonoides* is an outgroup to the parabasalian clade while *T. foetus* is an outgroup to *P. keilini,* and *T. vaginalis.*

Figure 3.39: Phylogeny of ATP binding/leucine-tRNA ligases/aminoacyl-tRNA ligase (Wlm17025 alignment). *P. keilini* is forming the longest branch on the tree and surprisingly and unlike the other tree topologies, it is not grouping with either *T. vaginalis,* or *T. foetus.* It is rather forming a clade with the free-living excavate *Bodo saltans* but with 41 bootstrap support only. They are also both branchng within other excavates on the tree.

Figure 3.40: Phylogeny of ribosomal protein S15A (Wlm17026 alignment). *P. keilini* is forming a clade with *T. vaginalis* , while *T. foetus* is an outgroup.

Figure 3.41: Phylogeny of Ribosomal protein S10p/S20e family protein (Wlm17027 alignment). Prabasalids are on the longest branch on the tree. *P. keilini* is in the same clade with *T. vaginalis*, while *T. foetus* is their outgroup with maximum bootstrap support.

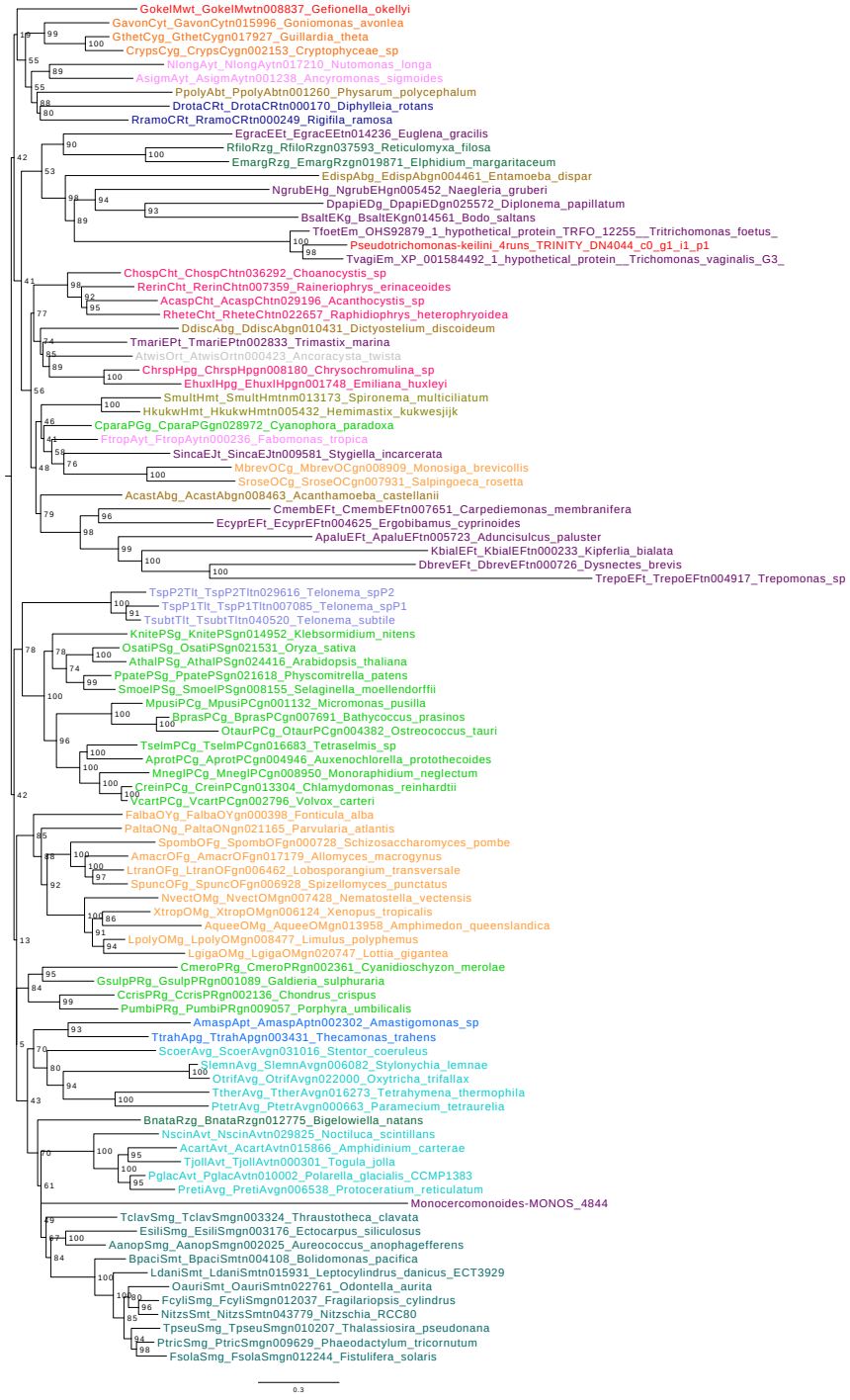

Figure 3.42: Phylogeny of R-protein L3 B (Wlm17028 alignment). The three parabasalids are grouping with other excavates on the tree including the free-living *Bodo saltans*. In the parabasalian clade, *P. keilini* and *T. vaginalis* are forming a clade for which *T. foetus* is an outgroup.

Figure 3.43: Phylogeny of Nucleotidylyl transferase superfamily protein (Wlm17029 alignment). The three parabasalids are grouping with other excavates on the tree. In the parabasalian clade, *P. keilini* and *T. vaginalis* are forming a clade for which *T. foetus* is an outgroup.

Figure 3.44: Phylogeny of Zinc-binding ribosomal protein family protein (Wlm17030 alignment). *P. keilini* is forming a clade with *T. foetus* with 61 bootstrap support, while *T. vaginalis* is their outgroup.

Figure 3.45: Phylogeny of Eukaryotic release factor 1 (eRF1) family protein (Wlm17031 alignment). Parabasalids are branching with other excavates like *Dysnectes brevis,* and *Carpediemonas membranifera.* In the parabasalian clade, *P. keilini* forms a clade with *T. vaginalis*, while *T. foetus* is their outgroup.

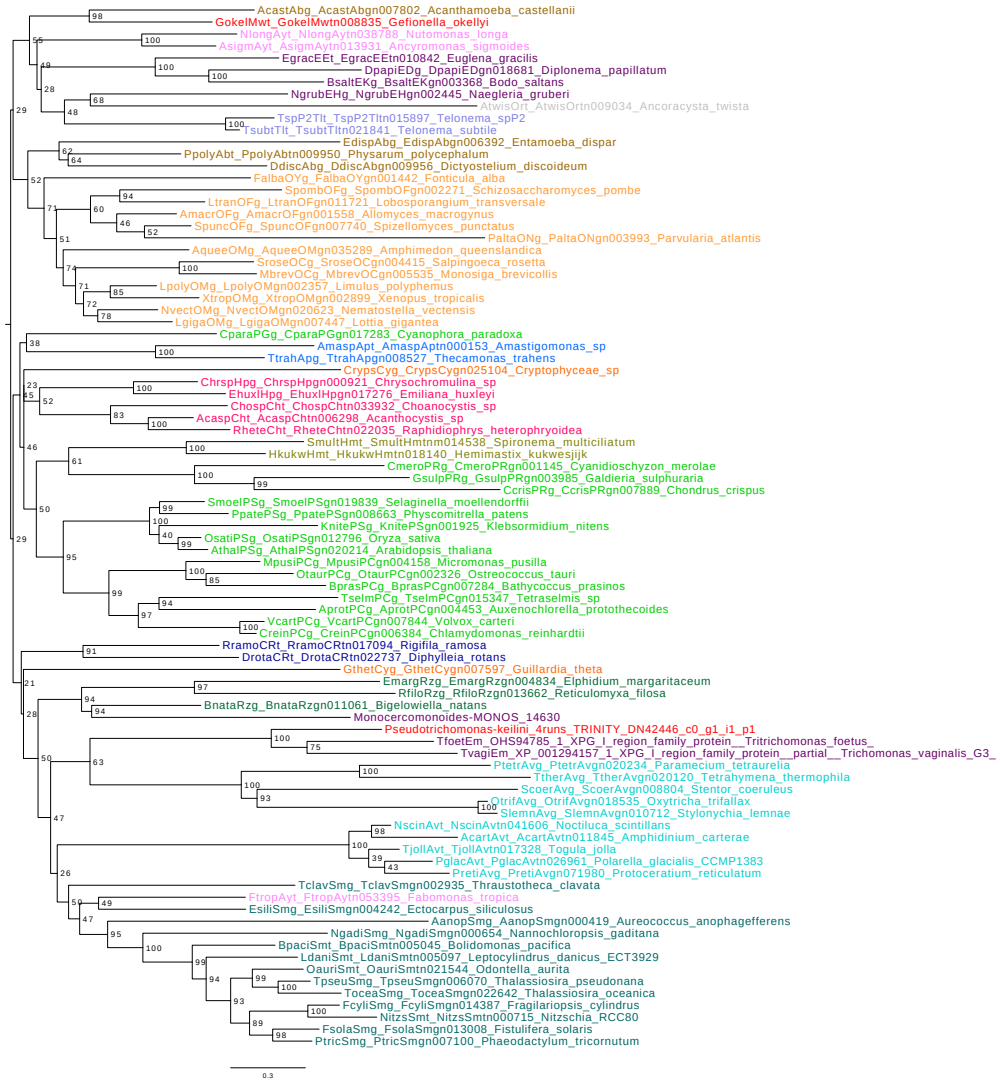

Figure 3.46: Phylogeny of 5'-3' exonuclease family protein (Wlm17032 alignment). *P. keilini* is the outgroup to the clade of Trichomonas composing of *T. vaginalis*, and *T. foetus*.

Figure 3.47: Phylogeny of Eukaryotic translation initiation factor 2 subunit 1 (Wlm17033 alignment). Parabasalids form the longest branch on the tree with *P. keilini* being the outgroup to the Trichomonas clade of *T. vaginalis,* and *T. foetus.*

Figure 3.48: Phylogeny of deoxyhypusine synthase (Wlm17034 alignment). *T. vaginalis* is an outgroup to the clade of *T. foetus*, and *P. keilini*.

Figure 3.49: Phylogeny of Actin-like ATPase superfamily protein (Wlm17035 alignment). *Monocercomonoides* is an outgroup to the parabasalian clade of *P. keilini, T. vaginalis,* and *T. foetus* and together the are are forming the longest branch on the tree.

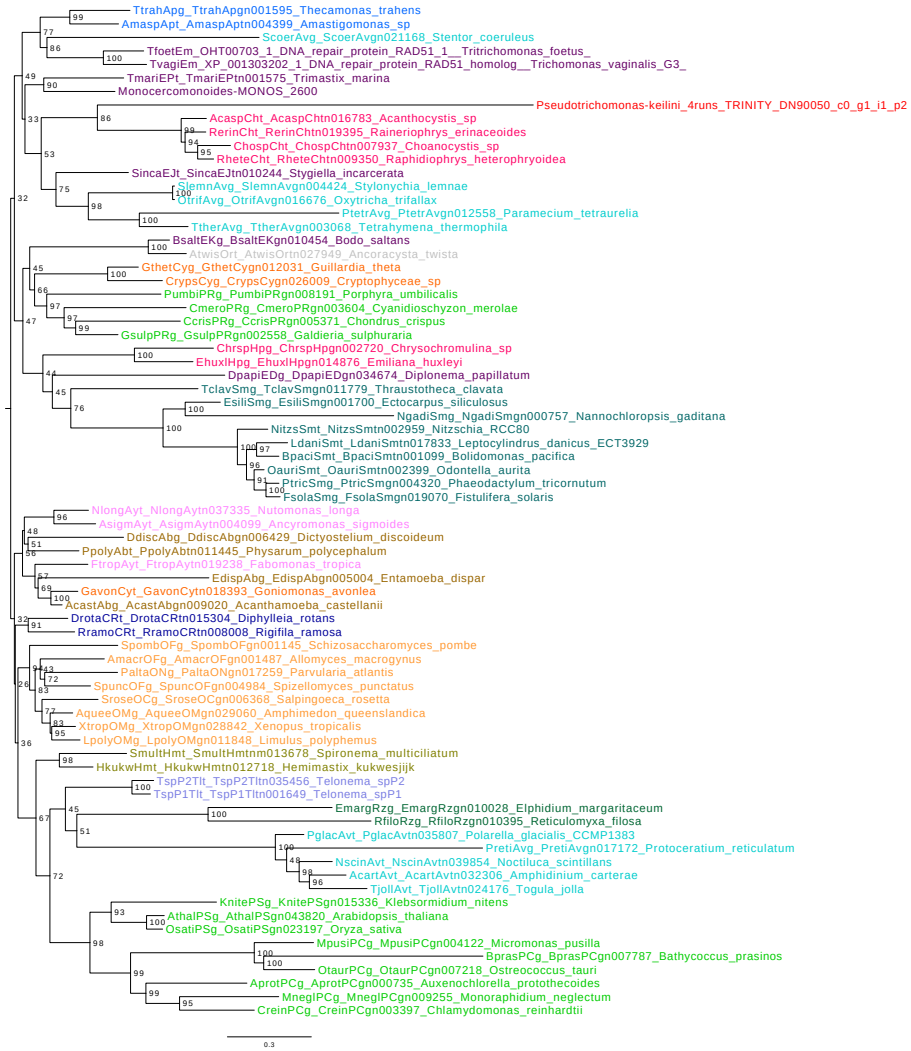

Figure 3.50: Phylogeny of DNA repair protein RAD51 homolog 1 (Wlm17036 alignment). Interestingly, *P. keilini* is not grouping with any of the parabasalids or excavates in the tree and is rather branching with Haptista and is forming the longest branch on the tree.

Figure 3.51: Phylogeny of Ribosomal protein S5 family protein (Wlm17037 alignment). In this tree, *P. keilini* is forming a clade with *T. vaginalis* with *T. foetus* being the outgroup.

Figure 3.52: Phylogeny of fibrillarin 2 (Wlm17038 alignment). *P. keilini* is nicely branching within nine other excavates in the tree. It is mainly forming a clade with *T. vaginalis* with maximum bootstrap support value, while *T. foetus* is their outgroup.

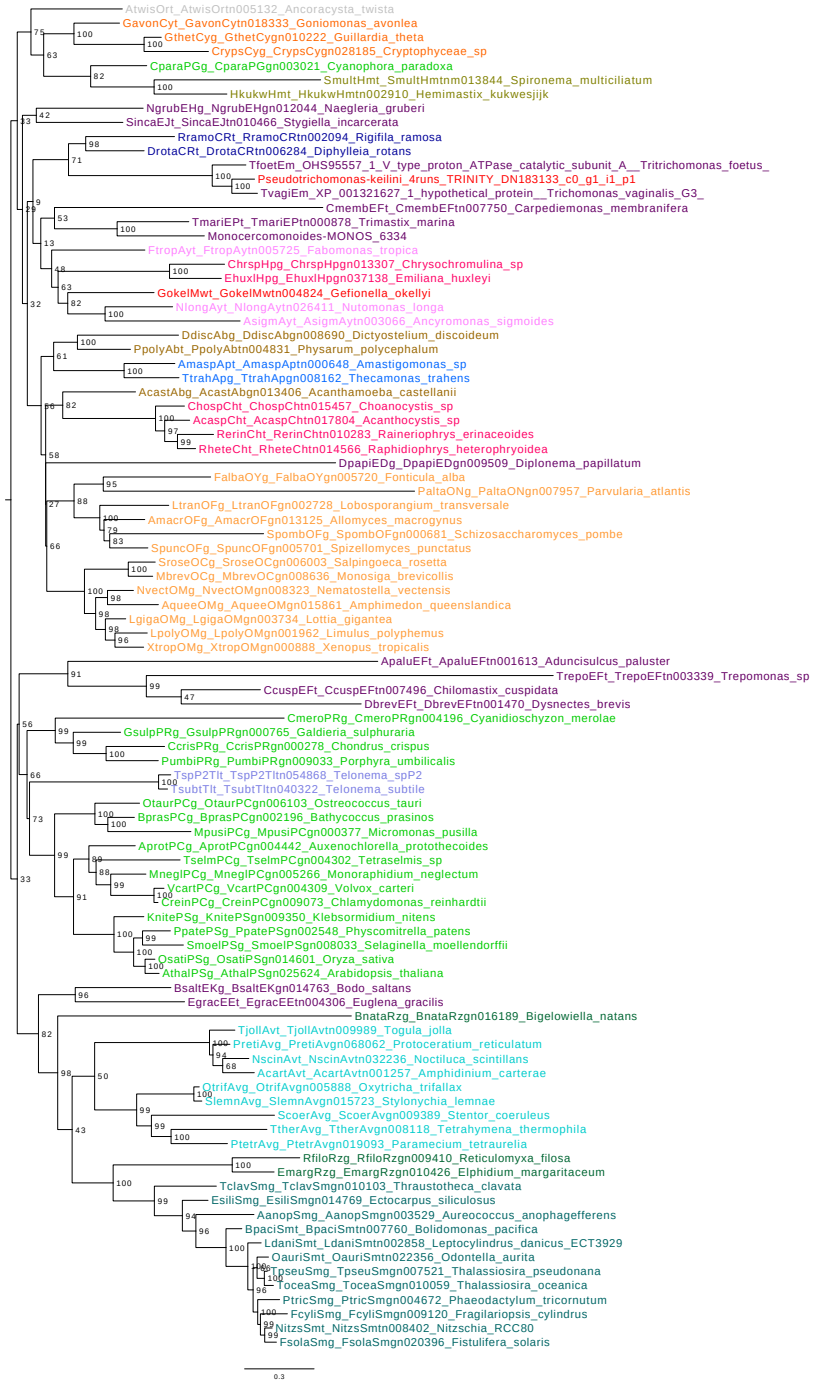

Figure 3.53: Phylogeny of vacuolar ATP synthase subunit A (Wlm17039 alignment). In this tree, *P. keilini* is forming a clade with *T. vaginalis* with *T. foetus* being their outgroup.

Figure 3.54: Phylogeny of Ribosomal protein S4 (RPS4A) family protein (Wlm17040 alignment). *P. keilini* is forming a clade with *T. vaginalis* with *T. foetus* being their outgroup.

Figure 3.55: Phylogeny of Ribosomal protein L2 family (Wlm17041 alignment). Parabasalids form the longest branch on the tree with *P. keilini* being the outgroup for the Trichomonads *T. vaginalis*, and *T. foetus*.

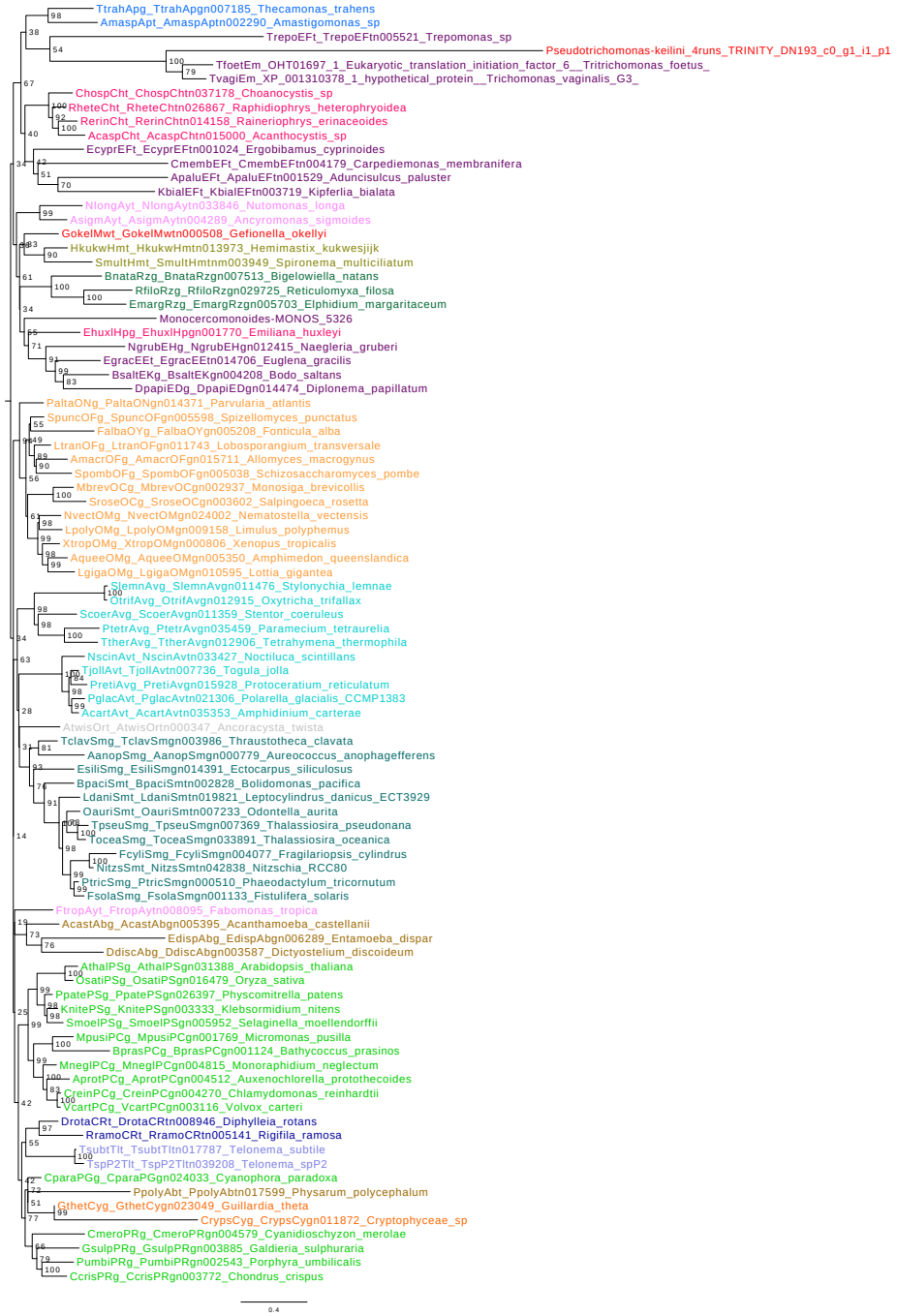

Figure 3.56: Phylogeny of Translation initiation factor IF6 (Wlm17042 alignment). *P. keilini* forms the longest branch on the tree and is the outgroup for the Trichomonads *T. vaginalis,* and *T. foetus.*, while *Trepomonas sp* is the outgroup to the parabasalian clade.

Figure 3.57: Phylogeny of Ribosomal protein S3 family protein (Wlm17044 alignment). In this tree, only two out of the three parabasalids are grouping together. *P. keilini* is forming a clade with *T. foetus*. Interestingly, *T. vaginalis* is on the longest branch on the tree.

Chapter 4

Conclusion

4.1 Transcriptomic data of *P. keilini*

Through this research, we present the first transcriptomic data for *Pseudotrichomonas keilini*, the first free-living parabasalid discovered in 1930s by Anne Bishop [24, 26] with only well-studied parasitic relatives. We did some analysis to try and discover its metabolic capabilities in comparison to the its closely related organisms *Trichomonas vaginalis* and *Tritrichomonas foetus*.

The size of the transcriptomic data is highly comparable with its relatives. The proteome of *P. keilini* contains 21,300 proteins which is close to the size of the proteome in *Tritrichomonas foetus*, which is 25,538. Since this is the first attempt to produce the transcriptomic data for *P. keilini*, there is a possiblity of not having the complete set of proteins. However, based on the metabolic analysis, we found key enzymes playing important roles in different metabolic pathways which was an indication to the completeness of the proteome. Another tool that was used to measure the completeness was BUSCO [168] which the uses conserved single-copy genes across eukaryotes as a reference for the completeness of the data. According to BUSCO, *P. keilini* contains 20 complete BUSCOs compared to *Trichomonas vaginalis* which has 22 BUSCOs.

However, it is important to remember that BUSCO results are not always a strong reference given that some of the *P. keilini* proteins may have not been detected.

4.2 The origin of the hydrogenenosome in parabasalids

Based on the analysis result of the hydrogenosomaal trees, as all the parabasalids enzymes were grouping together and with the other eukaryotic ones as well, we conclude that the common ancestor for parabasalids harboured a hydrogenosome.

This is in fact consistent with literature especially for PFO and [Fe-Fe] hydrogenase enzymes which were hypothesized to have been present in the early eukaryotes as it evolved from mitochondria that lost its ability for oxidatiove phosphorylation and gained the ability to produce hydrogen instead. We are yet to discover how that transition happened and how the organelle loses and gains functions. To fully understand the picture and the transitional stages in the evolution of hydrogenosomes we should be looking for an organism like *Nyctotherus* which already represents a transitional stage since its hydrogenosome retained its genome. Another support for the hypothesis that the common ancestor of parabasalids contained hydrogenosomes is based on the hypothesis that *Trichomonas* split early from other eukaryotes to live in anaerobic environments and never actually had mitochondria, and that hydrogenosomoes and the classic mitochondria shared a common progenitor instead [35, 52].

Our analysis of the key hydrogenosomal enzymes such as PFO and [Fe-Fe]hydrogenase agrees with other work that the eukaryotic sequences are generally monophyletic but their prokaryotic origin is unclear from the phylogenetic trees. While we do not know when exactly these enzymes were incorporated into the eukaryotic cell, we propose that they can be traced back to the common ancestor of parabasalids as they played a vital role in their survival through energy production.

4.3 Metabolism of *P. keilini*

Having few well-studied relatives, especially free-living ones, we mainly based our analyses on the available data for its parasitic relatives; *T. vaginalis,* and *T. foetus.* Since carbohydrates were discovered to be the preferred substrate for *T. vaginalis*, we conducted the analysis with the hypothesis that carbohydrates metabolism will be the preffered energy-production pathway for *P. keilini.* Our bioinformatics pipeline analysis has supported this hypothesis as we detected all the key enzymes involved in carbohydrates metabolism.

We further investigated the ability of *P. keilini* to metabolize lipids in comparison to its free-living relative *N. gruberi* which prefers lipids metabolism for energy production. Our analysis did not indicate the presence of the required enzymes in *P. keilini* for these biochemical pathways. This shows that the metabolism of *P. keilini* is more similar to its parasitic relative than that of its free-living one.

4.4 Phylogeny of excavates

According to our literature review, there is no clear consensus on whether excavates are monophyletic or not. In an attempt to resolve the interrelationships between the

different sub-clades of excavates, we conducted two kinds of phylogenetic analyses. The first, was through concatenating the orthologs of different metamonads and eukaryotic species, which did not support the monphyly of excavates. As for the second one, a supermatrix analysis that was done through the concatenation of 41 alignments of eukaryotic marker proteins that were identified by Williams and colleagues (2017). The tree resulting from the supermatrix analysis was supporting the monophyly of excavates. These contrasting results show that further work is yet to be done before we can conclude the monophyly of excavates.

4.4.1 Phylogenetics of *P. keilini*

In our analysis of a species tree which included sequences from metamonads and other eukaryotes, the phylogeny showed that *P. keilini* is grouping with parasitic relatives such as *T. vaginalis,* and *T. foetus* and not with other free-living ones. In most of the trees, *P. keilini* is forming a clade with *T. vaginalis* with *T. foetus* being the outgroup of their clade.

4.5 Further work

As this is the first genome-wide study of the genome and transcriptome of the free-living parabasalid *Pseudotrichomonas keilini*, it forms a preliminary exploratory study of the organism and its metabolic capabilities. Hence, further work needs to be done to better understand its biochemical capabilities and genomic structure.

The next steps in this project can address the following questions:

1. How the bacteriovorus *P. keilini* degrades the bacterial membranes which are composed of phospholipids bilayer, given the incomplete lipid metabolism pathway.

2. We also need to understand the origin and evolution of parasitism in *Trichomonas* by doing comparative genomics analyses of *P. keilini* against its parasitic relatives.

3. Structural genomics analyses on the genome of *P. keilini*.

4. Comparing the proteomics and metabolomics of the *P. keilini* hydrogenosome against that of *T. vaginalis* on a biochemical level, can also shed lights on the evolution of the organelle and the adaptation to the anaerobic environment.

Bibliography

[1] Ankur Abhishek, Anish Bavishi, Ashay Bavishi, and Madhusudan Choudhary. Bacterial genome chimaerism and the origin of mitochondria. *Canadian journal of microbiology*, 57(1):49–61, 2011. tex.publisher: NRC Research Press.

[2] Sina M Adl, David Bass, Christopher E Lane, Julius Lukeš, Conrad L Schoch, Alexey Smirnov, Sabine Agatha, Cédric Berney, Matthew W Brown, Fabien Burki, and others. Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1):4–119, 2019. Publisher: Wiley Online Library.

[3] Sina M Adl, Brian S Leander, Alastair GB Simpson, John M Archibald, O Roger Anderson, David Bass, Samuel S Bowser, Guy Brugerolle, Mark A Farmer, Sergey Karpov, and others. Diversity, nomenclature, and taxonomy of protists. *Systematic Biology*, 56(4):684–689, 2007. tex.publisher: Society of Systematic Zoology.

[4] Sina M. Adl, Alastair G. B. Simpson, Christopher E. Lane, Julius Lukeš, David Bass, Samuel S. Bowser, Matthew W. Brown, Fabien Burki, Micah Dunthorn, Vladimir Hampl, Aaron Heiss, Mona Hoppenrath, Enrique Lara, Line Le Gall, Denis H. Lynn, Hilary McManus, Edward A. D. Mitchell, Sharon E. Mozley-Stanridge, Laura W. Parfrey, Jan Pawlowski, Sonja Rueckert, Laura Shadwick, Lora Shadwick, Conrad L. Schoch, Alexey Smirnov, and Frederick W. Spiegel. The revised classification of eukaryotes. *The Journal of Eukaryotic Microbiology*, 59(5):429–493, September 2012.

[5] Anna Akhmanova, Frank Voncken, Theo van Alen, Angela van Hoek, Brigitte Boxma, Godfried Vogels, Marten Veenhuis, and Johannes HP Hackstein. A hydrogenosome with a genome. *Nature*, 396(6711):527–528, 1998. tex.publisher: Nature Publishing Group.

[6] M Akhtar and HA El-Obeid. Inactivation of serine transhydroxymethylase and threonine aldolase activities. *Biochimica et Biophysica Acta (BBA)-Enzymology*, 258(3):791–799, 1972. tex.publisher: Elsevier.

[7] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. Publisher: Elsevier.

[8] H. Amiri, O. Karlberg, and S.G.E. Andersson. Deep origin of plastid/parasite ATP/ADP translocases. *Journal of Molecular Evolution*, 56(2):137–150, 2003.

[9] Siv GE Andersson and Charles G Kurland. Origins of mitochondria and hydrogenosomes. *Current opinion in microbiology*, 2(5):535–541, 1999. tex.publisher: Elsevier.

[10] Siv GE Andersson, Alireza Zomorodipour, Jan O Andersson, Thomas Sicheritz-Pontén, U Cecilia M Alsmark, Raf M Podowski, A Kristina Näslund, Ann-Sofie Eriksson, Herbert H Winkler, and Charles G Kurland. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, 396(6707):133–140, 1998. tex.publisher: Nature Publishing Group.

[11] A. Atteia, A. Adrait, S. Brugiere, M. Tardif, R. van Lis, O. Deusch, T. Dagan, L. Kuhn, B. Gontero, W. Martin, J. Garin, J. Joyard, and N. Rolland. A Proteomic Survey of Chlamydomonas reinhardtii Mitochondria Sheds New Light on the Metabolic Plasticity of the Organelle and on the Nature of the -Proteobacterial Mitochondrial Ancestor. *Molecular Biology and Evolution*, 26(7):1533–1548, July 2009.

[12] Ariane Atteia, Robert Van Lis, Aloysius GM Tielens, and William F Martin. Anaerobic energy metabolism in unicellular photosynthetic eukaryotes. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1827(2):210–223, 2013. tex.publisher: Elsevier.

[13] Pantelis G Bagos, Theodore D Liakopoulos, Ioannis C Spyropoulos, and Stavros J Hamodrakas. PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic acids research*, 32(suppl_2):W400–W404, 2004. Publisher: Oxford University Press.

[14] Maria José Barberà, Iñaki Ruiz-Trillo, Jessica Leigh, Laura A Hug, and Andrew J Roger. The diversity of mitochondrion-related organelles amongst eukaryotic microbes. In *Origin of mitochondria and hydrogenosomes*, pages 239–275. Springer, 2007.

[15] RA Beinart, DJ Beaudoin, JM Bernhard, and VP Edgcomb. Insights into the metabolic functioning of a multipartner ciliate symbiosis from oxygen-depleted sediments. *Molecular ecology*, 27(8):1794–1807, 2018. tex.publisher: Wiley Online Library.

[16] A Bekker, HD Holland, P-L Wang, DIII Rumble, HJ Stein, JL Hannah, LL Coetzee, and NJ Beukes. Dating the rise of atmospheric oxygen. *Nature*, 427(6970):117–120, 2004. tex.publisher: Nature Publishing Group.

[17] Marlene Benchimol. The Mastigont System in Trichomonads. In W. de Souza, editor, *Structures and Organelles in Pathogenic Protists*, Microbiology Monographs, pages 1–26. Springer, Berlin, Heidelberg, 2010.

[18] Catherine Bernard, Alastair G. B. Simpson, and David J. Patterson. Some free-living flagellates (protista) from anoxic habitats. *Ophelia*, 52(2):113–142, May 2000.

[19] Joan M Bernhard, Kurt R Buck, Mark A Farmer, and Samuel S Bowser. The santa barbara basin is a symbiosis oasis. *Nature*, 403(6765):77–80, 2000. tex.publisher: Nature Publishing Group.

[20] Michiel L Bexkens, Verena Zimorski, Maarten J Sarink, Hans Wienk, Jos F Brouwers, Johan F De Jonckheere, William F Martin, Fred R Opperdoes, Jaap J van Hellemond, and Aloysius GM Tielens. Lipids are the preferred substrate of the protist Naegleria gruberi, relative of a human brain pathogen. *Cell reports*, 25(3):537–543, 2018. Publisher: Elsevier.

[21] Giancarlo A. Biagini, Bland J. Finlay, and David Lloyd. Evolution of the hydrogenosome. *FEMS Microbiology Letters*, 155(2):133–140, January 2006.

[22] Ann Bishop. The morphology and method of division of Trichomonas. *Parasitology*, 23(2):129–156, 1931. tex.publisher: Cambridge University Press.

[23] Ann Bishop. A note upon Trichomonas sanguisugae Alexeieff 1911. *Parasitology*, 24(1):140–142, 1932. tex.publisher: Cambridge University Press.

[24] Ann Bishop. Observations upon a " *Trichomonas* " from Pond Water. *Parasitology*, 27(2):246–256, May 1935.

[25] Ann Bishop. Further Observations upon a " *Trichomonas* " from Pond Water. *Parasitology*, 28(3):443–445, July 1936.

[26] Ann Bishop. A note upon the systematic position of " *Trichomonas* " *keilini* (Bishop, 1935). *Parasitology*, 31(4):469–472, December 1939.

[27] RL Blakley. A spectrophotometric study of the reaction catalysed by serine transhydroxymethylase. *Biochemical Journal*, 77(3):459, 1960. tex.publisher: Portland Press Ltd.

[28] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[29] L Bonen, RS Cunningham, MW Gray, and WF Doolittle. Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature. *Nucleic acids research*, 4(3):663–671, 1977. tex.publisher: Oxford University Press.

[30] Thomas Bourguignon, Nathan Lo, Stephen L Cameron, Jan Šobotník, Yoshinobu Hayashi, Shuji Shigenobu, Dai Watanabe, Yves Roisin, Toru Miura, and Theodore A Evans. The evolutionary history of termites as inferred from 66 mitochondrial genomes. *Molecular Biology and Evolution*, 32(2):406–421, 2014. tex.publisher: Society for Molecular Biology and Evolution.

[31] G. Brugerolle. Flagellar and cytoskeletal systems in amitochondrial flagellates: Archamoeba, Metamonada and Parabasala. *Protoplasma*, 164(1):70–90, February 1991.

[32] G Brugerolle and JJ Lee. Phylum parabasalia. *An illustrated guide to the protozoa*, 2:1196–1250, 2000. tex.publisher: Allen Press Lawrence, KS, USA.

[33] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[34] Elizabeth TN Bui and Patricia J Johnson. Identification and characterization of [Fe]-hydrogenases in the hydrogenosome of Trichomonas vaginalis. *Molecular and biochemical parasitology*, 76(1-2):305–310, 1996. tex.publisher: Elsevier.

[35] ET Bui, Peter J Bradley, and Patricia J Johnson. A common evolutionary origin for mitochondria and hydrogenosomes. *Proceedings of the National Academy of Sciences*, 93(18):9651–9656, 1996. tex.publisher: National Acad Sciences.

[36] Elena Bushmanova, Dmitry Antipov, Alla Lapidus, and Andrey D. Prjibelski. rnaSPAdes: *a de novo* transcriptome assembler and its application to RNA-Seq data. preprint, Bioinformatics, September 2018.

[37] Sarah E. Calvo and Vamsi K. Mootha. The Mitochondrial Proteome and Human Disease. *Annual Review of Genomics and Human Genetics*, 11(1):25–44, 2010.

[38] Jane M. Carlton, Robert P. Hirt, Joana C. Silva, Arthur L. Delcher, Michael Schatz, Qi Zhao, Jennifer R. Wortman, Shelby L. Bidwell, U. Cecilia M. Alsmark, Sébastien Besteiro, Thomas Sicheritz-Ponten, Christophe J. Noel, Joel B. Dacks, Peter G. Foster, Cedric Simillion, Yves Van de Peer, Diego Miranda-Saavedra, Geoffrey J. Barton, Gareth D. Westrop, Sylke Müller, Daniele Dessi, Pier Luigi Fiori, Qinghu Ren, Ian Paulsen, Hanbang Zhang, Felix D. Bastida-Corcuera, Augusto Simoes-Barbosa, Mark T. Brown, Richard D. Hayes, Mandira Mukherjee, Cheryl Y. Okumura, Rachel Schneider, Alias J. Smith, Stepanka Vanacova, Maria Villalvazo, Brian J. Haas, Mihaela Pertea, Tamara V. Feldblyum, Terry R. Utterback, Chung-Li Shu, Kazutoyo Osoegawa, Pieter J. de Jong, Ivan Hrdy, Lenka Horvathova, Zuzana Zubacova, Pavel Dolezal, Shehre-Banoo Malik, John M. Logsdon, Katrin Henze, Arti Gupta, Ching C. Wang, Rebecca L. Dunne, Jacqueline A. Upcroft, Peter Upcroft, Owen White, Steven L. Salzberg, Petrus Tang, Cheng-Hsun Chiu, Ying-Shiung Lee, T. Martin Embley, Graham H. Coombs, Jeremy C. Mottram, Jan Tachezy, Claire M. Fraser-Liggett, and Patricia J. Johnson. Draft Genome Sequence of the Sexually Transmitted Pathogen Trichomonas vaginalis. *Science (New York, N.Y.)*, 315(5809):207–212, January 2007.

[39] Thomas Cavalier-Smith. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International journal of systematic and evolutionary microbiology*, 52(2):297–354, 2002. tex.publisher: Microbiology Society.

[40] Thomas Cavalier-Smith. Origin of mitochondria by intracellular enslavement of a photosynthetic purple bacterium. *Proceedings of the Royal Society B: Biological Sciences*, 273(1596):1943–1952, August 2006.

[41] Thomas Cavalier-Smith. Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. *European Journal of Protistology*, 49(2):115–178, May 2013.

[42] Ivan Cepicka, Vladimír Hampl, and Jaroslav Kulda. Critical Taxonomic Revision of Parabasalids with Description of one New Genus and three New Species. *Protist*, 161(3):400–433, July 2010.

[43] J. O. Corliss. An interim utilitarian ('User-friendly') hierarchical classification and characterization of the protists. *Acta Protozoologica*, 1994.

[44] Alexis Criscuolo and Simonetta Gribaldo. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informa-

tive regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10(1):210, July 2010.

[45] Joel B Dacks, Mark C Field, Roger Buick, Laura Eme, Simonetta Gribaldo, Andrew J Roger, Céline Brochier-Armanet, and Damien P Devos. The changing view of eukaryogenesis–fossils, cells, lineages and how they all come together. *Journal of cell science*, 129(20):3695–3703, 2016. tex.publisher: The Company of Biologists Ltd.

[46] Roberto Danovaro, Antonio Dell'Anno, Antonio Pusceddu, Cristina Gambi, Iben Heiner, and Reinhardt Møbjerg Kristensen. The first metazoa living in permanently anoxic conditions. *BMC biology*, 8:30, April 2010.

[47] Yaacov Davidov, Dorothee Huchon, Susan F. Koval, and Edouard Jurkevitch. A new -proteobacterial clade of Bdellovibrio-like predators: implications for the mitochondrial endosymbiotic theory. *Environmental Microbiology*, 8(12):2179–2188, 2006.

[48] Yaacov Davidov and Edouard Jurkevitch. Predation between prokaryotes and the origin of eukaryotes. *BioEssays*, 31(7):748–757, 2009.

[49] Elizabeth A Davidson, Mark van der Giezen, David S Horner, T Martin Embley, and Christopher J Howe. An [Fe] hydrogenase from the anaerobic hydrogenosome-containing fungus Neocallimastix frontalis L2. *Gene*, 296(1-2):45–52, 2002. tex.publisher: Elsevier.

[50] Fabienne Dufernez, Richard L Walker, Christophe Noel, Stephanie Caby, Clea Mantini, PILAR DELGADO-VISCOGLIOSI, Moriya Ohkuma, Toshiaki Kudo, Monique Capron, Raymond J Pierce, and others. Morphological and molecular identification of non-Tritrichomonas foetus trichomonad protozoa from the bovine preputial cavity. *Journal of eukaryotic microbiology*, 54(2):161–168, 2007. tex.publisher: Wiley Online Library.

[51] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme J. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. 1998.

[52] Sabrina D Dyall and Patricia J Johnson. Origins of hydrogenosomes and mitochondria: evolution and organelle biogenesis. *Current opinion in microbiology*, 3(4):404–411, 2000. tex.publisher: Elsevier.

[53] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, October 1998.

[54] Virginia P Edgcomb, Edward R Leadbetter, William Bourland, David Beaudoin, and Joan Bernhard. Structured multiple endosymbiosis of bacteria and archaea in a ciliate from marine sulfidic sediments: a survival mechanism in low oxygen, sulfidic sediments? *Frontiers in microbiology*, 2:55, 2011. tex.publisher: Frontiers.

[55] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, January 2009.

[56] Martin Embley, Mark van der Giezen, David S Horner, Patricia L Dyal, and Peter Foster. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1429):191–203, 2003. tex.publisher: The Royal Society.

[57] T Martin Embley and Bland J Finlay. The use of small subunit rRNA sequences to unravel the relationships between anaerobic ciliates and their methanogen endosymbionts. *Microbiology*, 140(2):225–235, 1994. tex.publisher: Microbiology Society.

[58] T. Martin Embley and William Martin. A hydrogen-producing mitochondrion. *Nature*, 396(6711):517–519, December 1998.

[59] T Martin Embley, Mark van der Giezen, David Horner, Patricia Dyal, Samantha Bell, and Peter Foster. Hydrogenosomes, mitochondria and early eukaryotic evolution. *IUBMB life*, 55(7):387–395, 2003. tex.publisher: Wiley Online Library.

[60] TM Embley, DA Horner, and RP Hirt. Anaerobic eukaryote evolution: hydrogenosomes as biochemically modified mitochondria? *Trends in ecology & evolution*, 12(11):437–441, 1997. tex.publisher: Elsevier.

[61] Laura Eme, Anja Spang, Jonathan Lombard, Courtney W. Stairs, and Thijs J. G. Ettema. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12):711–723, December 2017.

[62] E Escomel. Sur la dysenterie a trichomonas a arequipa (perou). *Bull. Soc. path. exot.*, 6:120, 1913.

[63] Mark A. Farmer. Ultrastructure of Ditrichomonas honigbergii N. G., N. Sp. (Parabasalia) and Its Relationship to Amitochondrial Protists. *The Journal of Eukaryotic Microbiology*, 40(5):619–626, September 1993.

[64] Tom Fenchel and Catherine Bernard. Endosymbiotic purple non-sulphur bacteria in an anaerobic ciliated protozoon. *FEMS Microbiology Letters*, 110(1):21–25, June 1993.

[65] Tom Fenchel and Bland J Finlay. *Ecology and evolution in anoxic worlds*. Oxford; New York: Oxford University Press, 1995, 1995.

[66] Adolph J. Ferro, Annabella Barrett, and Stanley K. Shapiro. Kinetic properties and the effect of substrate analogues on 5-methylthioadenosine nucleosidase from Escherichia coli. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 438(2):487–494, July 1976.

[67] Lillian K Fritz-Laylin, Simon E Prochnik, Michael L Ginger, Joel B Dacks, Meredith L Carpenter, Mark C Field, Alan Kuo, Alex Paredez, Jarrod Chapman, Jonathan Pham, and others. The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell*, 140(5):631–642, 2010. tex.publisher: Elsevier.

[68] Motoji Fujioka. Purification and properties of serine hydroxymethylase from soluble and mitochondrial fractions of rabbit liver. *Biochimica et Biophysica Acta (BBA)-Enzymology*, 185(2):338–349, 1969. tex.publisher: Elsevier.

[69] K. Fujiwara, K. Okamura-Ikeda, and Y. Motokawa. Mechanism of the glycine cleavage reaction. Further characterization of the intermediate attached to H-protein and of the reaction catalyzed by T-protein. *The Journal of Biological Chemistry*, 259(17):10664–10668, September 1984.

[70] Kazuko Fujiwara, Kazuko Okamura, and Yutaro Motokawa. Hydrogen carrier protein from chicken liver: Purification, characterization, and role of its prosthetic group, lipoic acid, in the glycine cleavage reaction. *Archives of Biochemistry and Biophysics*, 197(2):454–462, October 1979.

[71] Toni Gabaldón and Martijn A Huynen. Reconstruction of the proto-mitochondrial metabolism. *Science*, 301(5633):609–609, 2003. tex.publisher: American Association for the Advancement of Science.

[72] Toni Gabaldón and Martijn A Huynen. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS computational biology*, 3(11), 2007. tex.publisher: Public Library of Science.

[73] Ryan M. R. Gawryluk, Kenneth A. Chisholm, Devanand M. Pinto, and Michael W. Gray. Compositional complexity of the mitochondrial proteome of a unicellular eukaryote (Acanthamoeba castellanii, supergroup Amoebozoa) rivals that of animals, fungi, and plants. *Journal of Proteomics*, 109:400–416, September 2014.

[74] Michael W Gray. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences*, 112(33):10133–10138, 2015. tex.publisher: National Acad Sciences.

[75] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, April 2013.

[76] Brian J. Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. LeDuc, Nir Friedman, and Aviv Regev. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols*, 8(8), August 2013.

[77] Vladimir Hampl, Laura Hug, Jessica W Leigh, Joel B Dacks, B Franz Lang, Alastair GB Simpson, and Andrew J Roger. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences*, 106(10):3859–3864, 2009. tex.publisher: National Acad Sciences.

[78] Vladimir Hampl, Courtney W Stairs, and Andrew J Roger. The tangled past of eukaryotic enzymes involved in anaerobic metabolism. *Mobile genetic elements*, 1(1):71–74, 2011. tex.publisher: Taylor & Francis.

[79] Vladimír Hampl, Ivan Cepicka, Jaroslav Flegr, Jan Tachezy, and Jaroslav Kulda. Morphological and molecular diversity of the monocercomonadid genera Monocercomonas, Hexamastix, and Honigbergiella gen. nov. *Protist*, 158(3):365–383, 2007. tex.publisher: Elsevier.

[80] Vladimír Hampl, David S Horner, Patricia Dyal, Jaroslav Kulda, Jaroslav Flegr, Peter G Foster, and T Martin Embley. Inference of the phylogenetic position of oxymonads based on nine genes: support for Metamonada and Excavata. *Molecular biology and evolution*, 22(12):2508–2518, 2005. tex.publisher: Oxford University Press.

[81] Vladimır Hampl, Ivan Cepicka, Jaroslav Flegr, Jan Tachezy, and Jaroslav Kulda. Critical analysis of the topology and rooting of the parabasalian 16S rRNA tree. *Molecular Phylogenetics and Evolution*, 32(3):711–723, September 2004.

[82] Aaron A Heiss, Martin Kolisko, Fleming Ekelund, Matthew W Brown, Andrew J Roger, and Alastair GB Simpson. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *Royal Society open science*, 5(4):171707, 2018. Publisher: The Royal Society Publishing.

[83] Koichi Hiraga and Goro Kikuchi. The mitochondrial glycine cleavage system. Functional association of glycine decarboxylase and aminomethyl carrier protein. *Journal of Biological Chemistry*, 255(24):11671–11676, 1980. tex.publisher: ASBMB.

[84] A Hollande and J Carruette-Valentin. Les atractophores, l'induction du fuseau et la division cellulaire chez les Hypermastigines, étude infrastructurale et révision systématique des Trichonymphines et des Spirotrichonymphines. *Protistologica*, 7:5–100, 1971.

[85] A Hollande and others. Le problème du centrosome et la cryptopleuromitose atractophorienne chez Lophomonas striata. 1972.

[86] A Hollande and J Valentin. Appareil de Golgi, pinocytose, lysosomes, mitochondries, bactéries symbiotiques, atractophores et pleuromitose chez les Hypermastigines du genre Joenia. *Protistologica*, 5(1):39–86, 1969.

[87] B. M. Honigberg. Evolutionary and Systematic Relationships in the Flagellate Order Trichomonadida Kirby*. *The Journal of Protozoology*, 10(1):20–63, 1963.

[88] BM Honigberg. Remarks upon trichomonad affinities of certain parasitic protozoa. In *Progress in protozoology, abstracts of papers read at the IVth int. Congress of protozoology (september 1973)*, 1973. tex.organization: Université of Clermont ClermontFerrand.

[89] David S Horner, Peter G Foster, and T Martin Embley. Iron hydrogenases and the evolution of anaerobic eukaryotes. *Molecular biology and evolution*, 17(11):1695–1709, 2000. tex.publisher: Oxford University Press.

[90] David S Horner, Burkhard Heil, Thomas Happe, and T Martin Embley. Iron hydrogenases–ancient enzymes in modern eukaryotes. *Trends in biochemical sciences*, 27(3):148–153, 2002. tex.publisher: Elsevier.

[91] David S Horner, Robert P Hirt, and T Martin Embley. A single eubacterial origin of eukaryotic pyruvate: ferredoxin oxidoreductase genes: implications for the evolution of anaerobic eukaryotes. *Molecular biology and evolution*, 16(9):1280–1291, 1999. tex.publisher: Oxford University Press.

[92] Ivan Hrdy, Robert P. Hirt, Pavel Dolezal, Lucie Bardonová, Peter G. Foster, Jan Tachezy, and T. Martin Embley. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, 432(7017):618–622, December 2004. Number: 7017 Publisher: Nature Publishing Group.

[93] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122, August 2017.

[94] Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(Database issue):D286–D293, January 2016.

[95] Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, Ivica Letunic, Thomas Rattei, Lars J. Jensen, Christian von Mering, and Peer Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, January 2019.

[96] Laura A Hug, Alexandra Stechmann, and Andrew J Roger. Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes. *Molecular biology and evolution*, 27(2):311–324, 2010. tex.publisher: Oxford University Press.

[97] Hiroyuki Imachi, Masaru K Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, Katsuyuki Uematsu, Tetsuro Ikuta, Motoo Ito, and others. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*, pages 1–7, 2020. tex.publisher: Nature Publishing Group.

[98] Emmanuelle J Javaux and Kevin Lepot. The Paleoproterozoic fossil record: implications for the evolution of the biosphere during Earth's middle-age. *Earth-Science Reviews*, 176:68–86, 2018. tex.publisher: Elsevier.

[99] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[100] Minoru Kanehisa, Yoko Sato, and Kanae Morishima. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, 428(4):726–731, February 2016.

[101] O. Karlberg, B. Canbäck, C.G. Kurland, and S.G.E. Andersson. The dual origin of the yeast mitochondrial proteome. *Yeast*, 17(3):170–187, 2000.

[102] Anna Karnkowska and Vladimír Hampl. The curious case of vanishing mitochondria. *Microbial Cell*, 3(10):491–494.

[103] Anna Karnkowska, Vojtěch Vacek, Zuzana Zubáčová, Sebastian C. Treitli, Romana Petrželková, Laura Eme, Lukáš Novák, Vojtěch Žárský, Lael D. Barlow, Emily K. Herman, Petr Soukal, Miluše Hroudová, Pavel Doležal, Courtney W. Stairs, Andrew J. Roger, Marek Eliáš, Joel B. Dacks, Čestmír Vlček, and Vladimír Hampl. A Eukaryote without a Mitochondrial Organelle. *Current Biology*, 26(10):1274–1284, May 2016.

[104] Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

[105] Laura A Katz and Jessica R Grant. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Systematic biology*, 64(3):406–415, 2015. tex.publisher: Oxford University Press.

[106] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546, May 2019.

[107] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017.

[108] Vera Kozjak, Nils Wiedemann, Dusanka Milenkovic, Christiane Lohaus, Helmut E Meyer, Bernard Guiard, Chris Meisinger, and Nikolaus Pfanner. An essential role of Sam50 in the protein sorting and assembly machinery of the mitochondrial outer membrane. *Journal of Biological Chemistry*, 278(49):48520–48523, 2003. tex.publisher: ASBMB.

[109] Alexander J Krause, Benjamin JW Mills, Shuang Zhang, Noah J Planavsky, Timothy M Lenton, and Simon W Poulton. Stepwise oxygenation of the Paleozoic atmosphere. *Nature communications*, 9(1):1–10, 2018. tex.publisher: Nature Publishing Group.

[110] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology*, 235(5):1501–1531, February 1994.

[111] Hidehiko Kumagai, Takatoshi Nagate, Hajime Yoshida, and Hideaki Yamada. Threonine aldolase from Candida humicola: II. Purification, crystallization and properties. *Biochimica et Biophysica Acta (BBA)-Enzymology*, 258(3):779–790, 1972. tex.publisher: Elsevier.

[112] CG Kurland and SGE Andersson. Origin and evolution of the mitochondrial proteome. *Microbiol. Mol. Biol. Rev.*, 64(4):786–820, 2000. tex.publisher: Am Soc Microbiol.

[113] G. Lavier. Sur un Trichomonadidé libre des eaux stagnantes. *Annales de Parasitologie Humaine et Comparée*, 14(4):359–368, 1936.

[114] Chun Pong Lee, Nicolas L. Taylor, and A. Harvey Millar. Recent Advances in the Composition and Heterogeneity of the Arabidopsis Mitochondrial Proteome. *Frontiers in Plant Science*, 4, 2013.

[115] Michelle M Leger, Laura Eme, Laura A Hug, and Andrew J Roger. Novel hydrogenosomes in the microaerophilic jakobid Stygiella incarcerata. *Molecular*

biology and evolution, 33(9):2318–2336, 2016. tex.publisher: Oxford University Press.

[116] Michelle M. Leger, Ryan M. R. Gawryluk, Michael W. Gray, and Andrew J. Roger. Evidence for a Hydrogenosomal-Type Anaerobic ATP Generation Pathway in Acanthamoeba castellanii. *PLoS ONE*, 8(9):e69532, September 2013.

[117] Timothy M Lenton, Tais W Dahl, Stuart J Daines, Benjamin JW Mills, Kazumi Ozaki, Matthew R Saltzman, and Philipp Porada. Earliest land plants created modern levels of atmospheric oxygen. *Proceedings of the National Academy of Sciences*, 113(35):9704–9709, 2016. tex.publisher: National Acad Sciences.

[118] William H Lewis, Anders E Lind, Kacper M Sendra, Henning Onsbring, Tom A Williams, Genoveva F Esteban, Robert P Hirt, Thijs J G Ettema, and T Martin Embley. Convergent Evolution of Hydrogenosomes from Mitochondria by Gene Transfer and Loss. *Molecular Biology and Evolution*, page msz239, October 2019.

[119] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[120] Roland Lill, Kerstin Diekert, Anita Kaut, Heike Lange, Winfried Pelzer, Corinna Prohl, and Gyula Kispal. The essential role of mitochondria in the biogenesis of cellular iron-sulfur proteins. *Biological chemistry*, 380(10):1157–1166, 1999. tex.publisher: Walter de Gruyter.

[121] D.G. Lindmark and M. Muller. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate Tritrichomonas foetus, and its role in pyruvate metabolism. *Journal of Biological Chemistry*, 248(22):7724–7728, 1973.

[122] Abhijith Makki, Petr Rada, Vojtěch Žárský, Sami Kereche, Lubomír Kováčik, Marian Novotn, Tobias Jores, Doron Rapaport, and Jan Tachezy. Triplet-pore structure of a highly divergent TOM complex of hydrogenosomes in Trichomonas vaginalis. *PLoS biology*, 17(1):e3000098, 2019. tex.publisher: Public Library of Science.

[123] W. Martin and M. Müller. The hydrogen hypothesis for the first eukaryote. *Nature*, 392(6671):37–41, 1998.

[124] William Martin and Miklós Müller. The hydrogen hypothesis for the first eukaryote. *Nature*, 392(6671):37–41, 1998. tex.publisher: Nature Publishing Group.

[125] William F Martin. Symbiogenesis, gradualism, and mitochondrial energy in eukaryote origin. *Periodicum biologorum*, 119(3):141–158, 2017. tex.publisher: Hrvatsko prirodoslovno društvo.

[126] William F Martin. Too much eukaryote LGT. *BioEssays*, 39(12):1700115, 2017. tex.publisher: Wiley Online Library.

[127] William F Martin, Sriram Garg, and Verena Zimorski. Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140330, 2015. tex.publisher: The Royal Society.

[128] John P McCutcheon. From microbiology to cell biology: when an intracellular bacterium becomes part of its host cell. *Current opinion in cell biology*, 41:132–136, 2016. tex.publisher: Elsevier.

[129] John P McCutcheon and Nancy A Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26, 2012. tex.publisher: Nature Publishing Group.

[130] Wes McKinney. Data Structures for Statistical Computing in Python. page 6, 2010.

[131] Camila Braz Menezes, Amanda Piccoli Frasson, and Tiana Tasca. Trichomoniasis-are we giving the deserved attention to the most common non-viral sexually transmitted disease worldwide? *Microbial cell*, 3(9):404, 2016. Publisher: Shared Science Publishers.

[132] Marek Mentel and William Martin. Energy metabolism among eukaryotic anaerobes in light of Proterozoic ocean chemistry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1504):2717–2729, 2008. tex.publisher: The Royal Society London.

[133] Constantin Mereschkowsky. Uber natur und ursprung der chromatophoren im pflanzenreiche. *Biologisches Centralblatt*, 25:293–604, 1905.

[134] Bernhard Misof, Shanlin Liu, Karen Meusemann, Ralph S Peters, Alexander Donath, Christoph Mayer, Paul B Frandsen, Jessica Ware, Tomáš Flouri, Rolf G Beutel, and others. Phylogenomics resolves the timing and pattern of

insect evolution. *Science*, 346(6210):763–767, 2014. tex.publisher: American Association for the Advancement of Science.

[135] David Moreira, Sophie von der Heyden, David Bass, Purificación López-García, Ema Chao, and Thomas Cavalier-Smith. Global eukaryote phylogeny: combined small-and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. *Molecular phylogenetics and evolution*, 44(1):255–266, 2007. Publisher: Elsevier.

[136] M. Muller, M. Mentel, J. J. van Hellemond, K. Henze, C. Woehle, S. B. Gould, R.-Y. Yu, M. van der Giezen, A. G. M. Tielens, and W. F. Martin. Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiology and Molecular Biology Reviews*, 76(2):444–495, June 2012.

[137] Ulrich Mühlenhoff and Roland Lill. Biogenesis of iron–sulfur proteins in eukaryotes: a novel task of mitochondria that is inherited from bacteria. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1459(2-3):370–382, 2000. tex.publisher: Elsevier.

[138] Miklós Müller. The hydrogenosome. *Microbiology*, 139(12):2879–2889, 1993. tex.publisher: Citeseer.

[139] Miklós Müller, Marek Mentel, Jaap J van Hellemond, Katrin Henze, Christian Woehle, Sven B Gould, Re-Young Yu, Mark van der Giezen, Aloysius GM Tielens, and William F Martin. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.*, 76(2):444–495, 2012. tex.publisher: Am Soc Microbiol.

[140] Natasha M Nesbitt, Camelia Baleanu-Gogonea, Robert M Cicchillo, Kathy Goodson, David F Iwig, John A Broadwater, Jeffrey A Haas, Brian G Fox, and Squire J Booker. Expression, purification, and physical characterization of Escherichia coli lipoyl (octanoyl) transferase. *Protein expression and purification*, 39(2):269–282, 2005. tex.publisher: Elsevier.

[141] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015. tex.publisher: Oxford University Press.

[142] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, January 2015.

[143] Julie EJ Nixon, Jessica Field, Andrew G McArthur, Mitchell L Sogin, Nigel Yarlett, Brendan J Loftus, and John Samuelson. Iron-dependent hydrogenases of Entamoeba histolytica and Giardia lamblia: activity of the recombinant entamoebic enzyme and evidence for lateral gene transfer. *The Biological Bulletin*, 204(1):1–9, 2003. tex.publisher: Marine Biological Laboratory.

[144] H. Nury, C. Dahout-Gonzalez, V. Trézéguet, G.J.M. Lauquin, G. Brandolin, and E. Pebay-Peyroula. Relations Between Structure and Function of the Mitochondrial ADP/ATP Carrier. *Annual Review of Biochemistry*, 75(1):713–741, June 2006.

[145] Aswini K. Panigrahi, Yuko Ogata, Alena Zíková, Atashi Anupama, Rachel A. Dalley, Nathalie Acestor, Peter J. Myler, and Kenneth D. Stuart. A comprehensive analysis of Trypanosoma brucei mitochondrial proteome. *PROTEOMICS*, 9(2):434–450, 2009.

[146] S. Pares, C. Cohen-Addad, L. Sieker, M. Neuburger, and R. Douce. X-ray structure determination at 2.6-A resolution of a lipoate-containing protein: the H-protein of the glycine decarboxylase complex from pea leaves. *Proceedings of the National Academy of Sciences*, 91(11):4850–4853, May 1994. Publisher: National Academy of Sciences Section: Research Article.

[147] Laura Wegener Parfrey, Daniel JG Lahr, Andrew H Knoll, and Laura A Katz. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences*, 108(33):13624–13629, 2011. tex.publisher: National Acad Sciences.

[148] Richard N Perham. Swinging arms and swinging domains in multifunctional enzymes: catalytic machines for multistep reactions. *Annual review of biochemistry*, 69(1):961–1004, 2000. tex.publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

[149] Dino Petrin, Kiera Delgaty, Renuka Bhatt, and Gary Garber. Clinical and microbiological aspects ofTrichomonas vaginalis. *Clinical microbiology reviews*, 11(2):300–317, 1998. tex.publisher: Am Soc Microbiol.

[150] Priscila Peña-Diaz and Julius Lukeš. Fe–S cluster assembly in the supergroup Excavata. *JBIC Journal of Biological Inorganic Chemistry*, 23(4):521–541, June 2018.

[151] Evelyn Plümper, Peter J Bradley, and Patricia J Johnson. Implications of protein import on the origin of hydrogenosomes. *Protist*, 149(4):303–311, 1998. tex.publisher: Elsevier.

[152] Susannah M Porter, Heda Agić, and Leigh Anne Riedman. Anoxic ecosystems and early eukaryotes. *Emerging Topics in Life Sciences*, 2(2):299–309, 2018. tex.publisher: Portland Press Ltd.

[153] Mascha Pusnik, Oliver Schmidt, Andrew J Perry, Silke Oeljeklaus, Moritz Niemann, Bettina Warscheid, Trevor Lithgow, Chris Meisinger, and André Schneider. Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin. *Current Biology*, 21(20):1738–1743, 2011. tex.publisher: Elsevier.

[154] Petr Rada, Pavel Doležal, Petr L Jedelsk, Dejan Bursac, Andrew J Perry, Miroslava Šedinová, Kateřina Smíšková, Marian Novotn, Neritza Campo Beltrán, Ivan Hrd, and others. The core components of organelle biogenesis and membrane transport in the hydrogenosomes of Trichomonas vaginalis. *PloS one*, 6(9), 2011. tex.publisher: Public Library of Science.

[155] Naiara Rodríguez-Ezpeleta and T Martin Embley. The SAR11 group of alphaproteobacteria is not related to the origin of mitochondria. *PloS one*, 7(1), 2012. tex.publisher: Public Library of Science.

[156] Andrew J Roger. Reconstructing early events in eukaryotic evolution. *the american naturalist*, 154(S4):S146–S163, 1999. tex.publisher: The University of Chicago Press.

[157] Andrew J. Roger, Sergio A. Muñoz-Gómez, and Ryoma Kamikawa. The Origin and Diversification of Mitochondria. *Current Biology*, 27(21):R1177–R1192, November 2017.

[158] Carmen Rotte, Frantisek Stejskal, Guan Zhu, Janet S Keithly, and William Martin. Pyruvate: NADP oxidoreductase from the mitochondrion of Euglena gracilis and from the apicomplexan Cryptosporidium parvum: a biochemical relic linking pyruvate metabolism in mitochondriate and amitochondriate protists. *Molecular Biology and Evolution*, 18(5):710–720, 2001. tex.publisher: Oxford University Press.

[159] Lynn Sagan. On the origin of mitosing cells. *Journal of Theoretical Biology*, 14(3):225–IN6, March 1967.

[160] Davide Sassera, Nathan Lo, Sara Epis, Giuseppe D'Auria, Matteo Montagna, Francesco Comandatore, David Horner, Juli Peretó, Alberto Maria Luciano, Federica Franciosi, and others. Phylogenomic evidence for the presence of a

flagellum and cbb 3 oxidase in the free-living mitochondrial ancestor. *Molecular biology and evolution*, 28(12):3285–3296, 2011. tex.publisher: Oxford University Press.

[161] Eric Sayers. *E-utilities Quick Start*. National Center for Biotechnology Information (US), October 2018.

[162] La Verne Schirch and Thomas Gross. Serine transhydroxymethylase identification as the threonine and allothreonine aldolases. *Journal of Biological Chemistry*, 243(21):5651–5655, 1968. tex.publisher: ASBMB.

[163] Andre Schneider. Mitochondrial protein import in trypanosomatids: Variations on a theme or fundamentally different? *PLoS pathogens*, 14(11), 2018. tex.publisher: Public Library of Science.

[164] Rachel E. Schneider, Mark T. Brown, April M. Shiflett, Sabrina D. Dyall, Richard D. Hayes, Yongming Xie, Joseph A. Loo, and Patricia J. Johnson. The Trichomonas vaginalis hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. *International Journal for Parasitology*, 41(13-14):1421–1434, November 2011.

[165] Robert M Schwartz and Margaret O Dayhoff. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science*, 199(4327):395–403, 1978. tex.publisher: JSTOR.

[166] D Searcy. Origins of mitochondria and chloroplasts from sulfurbased symbioses. *The origin and evolution of the cell*, pages 47–78, 1992.

[167] Alastair G. B. Simpson. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *International Journal of Systematic and Evolutionary Microbiology,*, 53(6):1759–1777, 2003.

[168] Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, 31(19):3210–3212, October 2015.

[169] Anja Spang, Jimmy H Saw, Steffen L Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs JG Ettema. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551):173–179, 2015. tex.publisher: Nature Publishing Group.

[170] Courtney W Stairs, Laura Eme, Matthew W Brown, Cornelis Mutsaers, Edward Susko, Graham Dellaire, Darren M Soanes, Mark Van Der Giezen, and Andrew J Roger. A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist Pygsuia. *Current Biology*, 24(11):1176–1186, 2014. tex.publisher: Elsevier.

[171] Courtney W. Stairs, Michelle M. Leger, and Andrew J. Roger. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140326, September 2015.

[172] Courtney W Stairs, Michelle M Leger, and Andrew J Roger. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140326, 2015. tex.publisher: The Royal Society.

[173] Courtney W Stairs, Andrew J Roger, and Vladimir Hampl. Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. *Molecular biology and evolution*, 28(7):2087–2099, 2011. tex.publisher: Oxford University Press.

[174] Alexandra Stechmann and Thomas Cavalier-Smith. The root of the eukaryote tree pinpointed. *Current Biology*, 13(17):R665–R666, 2003. Publisher: Elsevier.

[175] Alexander Steinbüchel and Miklós Müller. Anaerobic pyruvate metabolism of Tritrichomonas foetus and Trichomonas vaginalis hydrogenosomes. *Molecular and biochemical parasitology*, 20(1):57–65, 1986. tex.publisher: Elsevier.

[176] Thorsten Stoeck, Anke Behnke, Richard Christen, Linda Amaral-Zettler, Maria J Rodriguez-Mora, Andrei Chistoserdov, William Orsi, and Virginia P Edgcomb. Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC biology*, 7(1):72, 2009. tex.publisher: Springer.

[177] Daniel A Stolper and C Brenhin Keller. A record of deep-ocean dissolved O 2 from the oxidation state of iron in submarine basalts. *Nature*, 553(7688):323–327, 2018. tex.publisher: Nature Publishing Group.

[178] Radek Szklarczyk and Martijn A Huynen. Mosaic origin of the mitochondrial proteome. *Proteomics*, 10(22):4012–4024, 2010. tex.publisher: Wiley Online Library.

[179] SIGNHILD TAMM and SIDNEY L TAMM. The fine structure of the centriolar apparatus and associated structures in the flagellates Deltotrichonympha and Koruga. II. Division. *The Journal of Protozoology*, 20(2):245–252, 1973. tex.publisher: Wiley Online Library.

[180] Thorsten Thiergart, Giddy Landan, Marc Schenk, Tal Dagan, and William F Martin. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biology and Evolution*, 4(4):466–485, 2012. tex.publisher: Oxford University Press.

[181] Aloysius GM Tielens, Carmen Rotte, Jaap J van Hellemond, and William Martin. Mitochondria as we don't know them. *Trends in biochemical sciences*, 27(11):564–572, 2002. tex.publisher: Elsevier.

[182] Mark van der Giezen, Graeme M Birdsey, David S Horner, John Lucocq, Patricia L Dyal, Marlene Benchimol, Christopher J Danpure, and T Martin Embley. Fungal hydrogenosomes contain mitochondrial heat-shock proteins. *Molecular biology and evolution*, 20(7):1051–1061, 2003. tex.publisher: Oxford University Press.

[183] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1 0 Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv:1907.10121 [physics]*, July 2019. arXiv: 1907.10121.

[184] Frank GJ Voncken, Brigitte Boxma, Angela HAM van Hoek, Anna S Akhmanova, Godfried D Vogels, Martijn Huynen, Marten Veenhuis, and Johannes HP Hackstein. A hydrogenosomal [Fe]-hydrogenase from the anaerobic chytrid Neocallimastix sp. L2. *Gene*, 284(1-2):103–112, 2002. tex.publisher: Elsevier.

[185] Zhang Wang and Martin Wu. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PloS one*, 9(10), 2014. tex.publisher: Public Library of Science.

[186] Zhang Wang and Martin Wu. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific Reports*, 5:7949, 2015. tex.publisher: Nature Publishing Group.

[187] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, and others. Database resources of the national center for biotechnology information. *Nucleic acids research*, 34(suppl_1):D173–D180, 2006. Publisher: Oxford University Press.

[188] Kelly P Williams, Bruno W Sobral, and Allan W Dickerman. A robust species tree for the alphaproteobacteria. *Journal of bacteriology*, 189(13):4578–4586, 2007. tex.publisher: Am Soc Microbiol.

[189] Tom A. Williams, Gergely J. Szöllősi, Anja Spang, Peter G. Foster, Sarah E. Heaps, Bastien Boussau, Thijs J. G. Ettema, and T. Martin Embley. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*, 114(23):E4602–E4611, June 2017.

[190] C.R. Woese. Endosymbionts and mitochondrial origins. *Journal of Molecular Evolution*, 10(2):93–96, 1977.

[191] Michael A Yamin and YAMIN MA. Flagellates of the orders trichomonadida kirby, oxymonadida grassé, and hypermastigida grassi and foà reported from lower termites (isoptera families mastotermitidae, kalotermitidae, hodotermitidae, termopsidae, rhinotermitidae, and serritermitidae) and from the woodfeeding roach cryptocercus (dictyoptera: Cryptocercidae). 1979.

[192] Dan Yang, Y Oyaizu, H Oyaizu, Gary J Olsen, and Carl R Woese. Mitochondrial origins. *Proceedings of the National Academy of Sciences*, 82(13):4443–4447, 1985. tex.publisher: National Acad Sciences.

[193] Naoji Yubuki, Vít Céza, Ivan Cepicka, Akinori Yabuki, Yuji Inagaki, Takeshi Nakayama, Isao Inouye, and Brian S. Leander. Cryptic Diversity of Free-Living Parabasalids, Pseudotrichomonas keilini and Lacusteria cypriaca n. g., n. sp., as Inferred from Small Subunit rDNA Sequences: CRYPTIC DIVERSITY OF FREE-LIVING PARABASALIDS. *Journal of Eukaryotic Microbiology*, 57(6):554–561, November 2010.

[194] Katarzyna Zaremba-Niedzwiedzka, Eva F Caceres, Jimmy H Saw, Disa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W Seitz, Karthik Anantharaman, Piotr Starnawski, Kasper U Kjeldsen, and others. Asgard archaea

illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637):353–358, 2017. tex.publisher: Nature Publishing Group.

[195] Qianqian Zhang, Petr Táborsk, Jeffrey D Silberman, Tomáš Pánek, Ivan Čepička, and Alastair GB Simpson. Marine isolates of Trimastix marina form a plesiomorphic deep-branching lineage within Preaxostyla, separate from other known trimastigids (Paratrimastix n. gen.). *Protist*, 166(4):468–491, 2015. tex.publisher: Elsevier.

[196] Verena Zimorski, Chuan Ku, William F Martin, and Sven B Gould. Endosymbiotic theory for organelle origins. *Current opinion in microbiology*, 22:38–48, 2014. tex.publisher: Elsevier.

[197] Ivan Čepička, Michael F. Dolan, and Gillian H. Gile. Parabasalia. In John M. Archibald, Alastair G.B. Simpson, Claudio H. Slamovits, Lynn Margulis, Michael Melkonian, David J. Chapman, and John O. Corliss, editors, *Handbook of the Protists*, pages 1–44. Springer International Publishing, Cham, 2016.

[198] Ondřej Šmíd, Anna Matušková, Simon R Harris, Tomáš Kučera, Marián Novotn, Lenka Horvathova, Ivan Hrd, Eva Kutějová, Robert P Hirt, T Martin Embley, and others. Reductive evolution of the mitochondrial processing peptidases of the unicellular parasites Trichomonas vaginalis and Giardia intestinalis. *PLoS pathogens*, 4(12), 2008. tex.publisher: Public Library of Science.

Appendices

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants

Species name	Domain	Taxonomy
Cyanophora paradoxa		
Chondrus crispus		
Galdieria sulphuraria		
Cyanidioschyzon merolae		
Porphyra umbilicalis		
Chlamydomonas reinhardtii		
Volvox carteri		
Tetraselmis sp		
Monoraphidium neglectum		
Bathycoccus prasinos		**Archaeplastida**
Ostreococcus tauri		
Micromonas pusilla		
Auxenochlorella protothecoides		
Oryza sativa		
Arabidopsis thaliana		
Physcomitrella patens		
Selaginella moellendorffii		
Klebsormidium nitens		
Carpediemonas membranifera		
Dysnectes brevis		
Ergobibamus cyprinoides		
Kipferlia bialata		
Trepomonas sp		
Aduncisulcus paluster		
Chilomastix cuspidata		
Trimastix marina		
Naegleria gruberi		
Stygiella incarcerata		
Bodo saltans		
Diplonema papillatum		
Euglena gracilis		
Giardia lamblia ATCC50803		
Leishmania braziliensis		
Leishmania donovani species complex		
Leishmania infantum JPCM5		
Leishmania major Friedlin		
Leishmania mexicana MHOM GT 2001 U1103		**Excavates**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Leishmania panamensis	Eukaryotes	
Leptomonas pyrrhocoris		
Leptomonas seymouri		
Phytomonas sp isolateEM1		
Trypanosoma brucei brucei		
Trypanosoma brucei gambiens		
Trypanosoma congolense IL3000		
Trypanosoma cruzi CLBrener		
Trypanosoma cruzi Dm28c		
Trypanosoma cruzi ID25		
Trypanosoma cruzi marinkellei		
Trypanosoma equiperdum		
Trypanosoma grayi		
Trypanosoma rangeli SC58		
Trypanosoma theileri		
Trypanosoma vivax Y486		
Trichomonas vaginalis G3		
Tritrichomonas foetus		
Monocercomonoides		
Ectocarpus siliculosus		SAR
Thalassiosira pseudonana		
Fragilariopsis cylindrus		
Nannochloropsis gaditana		
Thraustotheca clavata		
Phaeodactylum tricornutum		
Thalassiosira oceanica		
Aureococcus anophagefferens		
Leptocylindrus danicus		
Odontella aurita		
Nitzschia RCC80		
Bolidomonas pacifica		
Fistulifera solaris		
Tetrahymena thermophila		
Paramecium tetraurelia		
Stylonychia lemnae		
Protoceratium reticulatum		
Noctiluca scintillans		
Togula jolla		
Polarella glacialis		

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Amphidinium carterae		
Oxytricha trifallax		
Stentor coeruleus		
Bigelowiella natans		
Reticulomyxa filosa		
Elphidium margaritaceum		
Cryptophyceae sp		**Cryptista**
Guillardia theta		
Goniomonas avonlea		
Emiliana huxleyi		**Haptista**
Chrysochromulina sp		
Raphidiophrys heterophryoidea		
Acanthocystis sp		
Choanocystis sp		
Raineriophrys erinaceoides		
Allomyces macrogynus		**Obazoa**
Schizosaccharomyces pombe		
Spizellomyces punctatus		
Lobosporangium transversale		
Parvularia atlantis		
Fonticula alba		
Amphimedon queenslandica		
Limulus polyphemus		
Xenopus tropicalis		
Nematostella vectensis		
Lottia gigantea		
Monosiga brevicollis		
Salpingoeca rosetta		
Gefionella okellyi		**Malawimonadidae**
Ancoracysta twista		**Janouskovec**
Ancyromonas sigmoides		**Eukaryota incertae sedis; Ancoracysta**
Fabomonas tropica		**Ancyromonadida; Planomonadidae; Fabomonas**
Nutomonas longa		**Ancyromonadida; Nutomonas**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Diphylleia rotans		**CRuMs; Collodictyonidae; Diphylleia**
Rigifila ramosa		**CRuMs; Rigifilida; Rigifila**
Candidatus Berkelbacteria bacterium GWA2_46_7		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium GWA2_35_9		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium GWA2_38_9		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium GWAE1_39_12		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium CG1_02_42_45		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium CG2_30_39_44		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium CG2_30_43_20		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium RIFCSPLOWO2_01_FULL_50_28		**CPR Berkelbacteria**
Candidatus Berkelbacteria bacterium RIFOXYA2_FULL_43_10		**CPR Berkelbacteria**
Candidatus Candidate division WS6 bacterium GW2011_GWA2_37_6		**CPR Dojkabacteria**
Candidate division WS6 bacterium GW2011_GWF1_35_23		**CPR Dojkabacteria**
Candidate division Kazan bacterium GW2011_GWA1_50_15		**CPR Kazan**
Candidate division Kazan bacterium GW2011_GWA1_44_22		**CPR Kazan**
Candidatus Amesbacteria bacterium RIFOXYB1_FULL_44_23		**CPR Microgenomates**
Candidatus Beckwithbacteria bacterium GW2011_GWC1_49_16		**CPR Microgenomates**
Candidatus Collierbacteria bacterium RIFOXYB1_FULL_49_13		**CPR Microgenomates**
Candidatus Curtissbacteria bacterium RBG_16_39_7		**CPR Microgenomates**
Candidatus Daviesbacteria bacterium RIFCSPHIGHO2_12_FULL_37_16		**CPR Microgenomates**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Candidatus Gottesmanbacteria bacterium GW2011_GWA1_43_11		**CPR Microgenomates**
Microgenomates group bacterium GW2011_GWA2_46_16		**CPR Microgenomates**
Candidatus Woykebacteria bacterium GWB1_45_5		**CPR Microgenomates**
Candidatus Levybacteria bacterium RIFCSPHIGHO2_02_FULL_37_10		**CPR Microgenomates**
Candidatus Pacebacteria bacterium CG1_02_43_31		**CPR Microgenomates**
Candidatus Woykebacteria bacterium RBG_13_40_15		**CPR Microgenomates**
Candidatus Woykebacteria bacterium RBG_13_40_7b		**CPR Microgenomates**
Candidatus Roizmanbacteria bacterium RIFCSPLOWO2_01_FULL_45_11		**CPR Microgenomates**
Candidatus Shapirobacteria bacterium GW2011_GWF2_37_20		**CPR Microgenomates**
Candidatus Woesebacteria bacterium GW2011_GWD2_40_19		**CPR Microgenomates**
Candidatus Woesebacteria bacterium RIFCSPHIGHO2_01_FULL_41_10		**CPR Microgenomates**
Candidatus Falkowbacteria bacterium GW2011_GWE2_38_254		**CPR Parcubacteria**
Candidatus Giovannonibacteria bacterium RIFCSPHIGHO2_02_43_16		**CPR Parcubacteria**
Candidatus Jorgensenbacteria bacterium GWA1_54_12		**CPR Parcubacteria**
Candidatus Jorgensenbacteria bacterium RIFCSPHIGHO2_02_FULL_45_20		**CPR Parcubacteria**
Candidatus Kaiserbacteria bacterium RIFCSPHIGHO2_01_FULL_56_24		**CPR Parcubacteria**
Parcubacteria bacterium SCGC AAA011-A09		**CPR Parcubacteria**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Candidatus Magasanikbacteria bacterium GW2011_GWA2_45_39		**CPR Parcubacteria**
Candidatus Magasanikbacteria bacterium GW2011_GWA2_46_17		**CPR Parcubacteria**
Candidatus Moranbacteria bacterium RIFCSPHIGHO2_01_FULL_55_24		**CPR Parcubacteria**
Candidatus Moranbacteria bacterium GW2011_GWC2_37_8		**CPR Parcubacteria**
Candidatus Nomurabacteria bacterium CG1_02_31_12		**CPR Parcubacteria**
Candidatus Nomurabacteria bacterium RIFCSPLOWO2_01_FULL_36_10b		**CPR Parcubacteria**
Candidatus Uhrbacteria bacterium RIFCSPLOWO2_02_FULL_49_11		**CPR Parcubacteria**
Candidatus Uhrbacteria bacterium RIFCSPLOWO2_02_FULL_54_37		**CPR Parcubacteria**
Candidatus Uhrbacteria bacterium RIFOXYC2_FULL_47_19		**CPR Parcubacteria**
Candidatus Wolfebacteria bacterium RIFOXYB1_FULL_54_12		**CPR Parcubacteria**
Candidatus Yanofskybacteria bacterium RIFCSPHIGHO2_01_FULL_44_17		**CPR Parcubacteria**
Candidatus Peregrinibacteria bacterium CG1_02_41_10		**CPR Peregrinibacteria**
Candidatus Peregrinibacteria bacterium CG1_02_54_53		**CPR Peregrinibacteria**
Candidatus Saccharibacteria bacterium CG2_30_41_52		**CPR Saccharibacteria**
Candidatus Saccharibacteria bacterium RIFCSPHIGHO2_12_FULL_49_19		**CPR Saccharibacteria**
Candidate division WWE3 bacterium RIFCSPLOWO2_01_FULL_42_11		**CPR WWE3**
Chloroherpeton thalassium ATCC 35110		**FBC Chlorobi**
Gemmatimonadetes bacterium RIFCSPLOWO2_12_FULL_68_9		**FBC Gemmatimonadetes**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Ignavibacteria bacterium GWA2_55_25	Bacteria	**FBC Ignavibacteria**
Candidate division Zixibacteria bacterium SM23_73_2		**FBC Zixibacteria**
Thermithiobacillus tepidarius DSM 3134		**Proteobacteria Acidithiobacillia**
Kordiimonas gwangyangensis DSM 19435		**Proteobacteria AlphaProteobacteria**
Pelagibacterium halotolerans B2		**Proteobacteria AlphaProteobacteria**
Rhodobacteraceae bacterium CG2_30_10_405		**Proteobacteria AlphaProteobacteria**
Granulibacter bethesdensis CGDNIH1		**Proteobacteria AlphaProteobacteria**
Tistrella mobilis KA081020 65		**Proteobacteria AlphaProteobacteria**
Anaplasma marginale		**Proteobacteria AlphaProteobacteria**
Citromicrobium sp.JLT1363		**Proteobacteria AlphaProteobacteria**
Bordetella bronchiseptica RB50		**Proteobacteria BetaProteobacteria**
Comamonas testosteroniTK102		**Proteobacteria BetaProteobacteria**
Gallionella capsiferriformans ES 2		**Proteobacteria BetaProteobacteria**
Methylovorus sp. SIP3 4		**Proteobacteria BetaProteobacteria**
Deefgea rivuli DSM 18356		**Proteobacteria BetaProteobacteria**
Bacteriovorax marinus SJ		**Proteobacteria DeltaProteobacteria**
Desulfotignum phosphitoxidansFiPS 3		**Proteobacteria DeltaProteobacteria**
Myxococcusxanthus DK 1622		**Proteobacteria DeltaProteobacteria**
Syntrophus aciditrophicus SB		**Proteobacteria DeltaProteobacteria**
Lebetimonas sp. JS032		**Proteobacteria EpsilonProteobacteria**
Anaerobiospirillum succiniciproducens DSM 6400		**Proteobacteria GammaProteobacteria**
Haliea rubraCM4115a DSM		**Proteobacteria GammaProteobacteria**
Arhodomonas aquaeolei DSM 8974		**Proteobacteria GammaProteobacteria**
Arsenophonus nasoniae DSM 15247		**Proteobacteria GammaProteobacteria**
Legionella micdadei ATCC 33218		**Proteobacteria GammaProteobacteria**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Methyloglobulus morosus KoM1		**Proteobacteria GammaProteobacteria**
Saccharospirillum impatiens DSM 12546		**Proteobacteria GammaProteobacteria**
Alkanindiges illinoisensis DSM 15370		**Proteobacteria GammaProteobacteria**
Photobacterium profundum SS9		**Proteobacteria GammaProteobacteria**
Frateuria terrea CGMCC 1.7053		**Proteobacteria GammaProteobacteria**
Mariprofundus ferrooxydans M34		**Proteobacteria ZetaProteobacteria**
RIFCSPLOWO2 2 FULL 45 22		**PVC Chlamydiae**
Chlamydophila abortus S263		**PVC Chlamydiae**
Lentisphaerae bacterium GWF2_52_8		**PVC Lentisphaerae**
Candidatus Omnitrophica bacterium CG1_02_46_14		**PVC Omnitrophica**
Planctomycetes bacterium GWB2_41_19		**PVC Planctomycetes**
Phycisphaerae bacterium SM1_79		**PVC Plantomycetes**
Arcanobacterium haemolyticum DSM 20595		**Terrabcteria Actinobacteria**
Corynebacterium argentoratense DSM 44202		**Terrabcteria Actinobacteria**
Gulosibactermolinativorax DSM 13485		**Terrabcteria Actinobacteria**
Aestuariimicrobium kwangyangense DSM 21549		**Terrabcteria Actinobacteria**
Bifidobacterium animalis animalis ATCC 25527		**Terrabcteria Actinobacteria**
Eggerthella sp. YY7918		**Terrabcteria Actinobacteria**
Rhodoluna lacicola MWH Ta8		**Terrabcteria Actinobacteria**
Enorma massiliensis phI		**Terrabcteria Actinobacteria**
Anaerolinea thermophila UNI-1		**Terrabcteria Chloroflexi**
Dehalococcoidia bacterium DG_22		**Terrabcteria Chloroflexi**
Chloroflexi bacterium RBG_13_50_10		**Terrabcteria Chloroflexi**
Aphanizomenon flos aquae NIES 81		**Terrabcteria Cyanobacteria**
Crinalium epipsammum PCC 9333		**Terrabcteria Cyanobacteria**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Oscillatoria sp. PCC 7112		**Terrabcteria Cyanobacteria**
Stanieria cyanosphaera PCC 7437		**Terrabcteria Cyanobacteria**
Bacillus anthracis 52 G		**Terrabcteria Firmicutes**
Marinococcus halotolerans DSM 16375		**Terrabcteria Firmicutes**
Fructobacillus fructosus KCTC 3544		**Terrabcteria Firmicutes**
Streptococcus mutans GS 5		**Terrabcteria Firmicutes**
Youngiibacter fragile 232.1		**Terrabcteria Firmicutes**
Filifactor alocis ATCC 35896		**Terrabcteria Firmicutes**
Orenia marismortui DSM 5156		**Terrabcteria Firmicutes**
Veillonella parvula DSM 2008		**Terrabcteria Firmicutes**
Melainabacteria bacterium MEL_A1		**Terrabcteria Melainabacteria**
Candidatus Melainabacteria bacterium RIFCSPLOWO2_12_FULL_35_11		**Terrabcteria Melainabacteria**
Mesoplasma florum W37		**Terrabcteria Tenericutes**
Thermosynechococcus elongatus BP1		**Terrabcteria Cyanobacteria**
Deinococcusgeothermalis DSM 11300		**Terrabcteria Deinococcus-Thermus**
Acidobacteria bacterium RBG_16_68_9		**Acidobacteria**
Acidobacteria bacterium RIFCSPLOWO2_02_FULL_68_18		**Acidobacteria**
Acidobacterium sp. MP5ACTX8		**Acidobacteria**
Holophaga foetida TMBS4 DSM 6591		**Acidobacteria**
Candidatus Aminicenantes bacterium RBG_16_66_30		**Aminicenantes**
Hydrogenobacter thermophilus TK 6		**Aquificae**
Thermovibrio ammonificans HB 1		**Aquificae**
Armatimonadetes bacterium 13_1_40CM_64_14		**Armatimonadetes**
Bacteroides fragilis NCTC 9343		**Bacteroidetes**
Marinilabilia salmonicolor JCM 21150		**Bacteroidetes**
Cyclobacterium marinum DSM 745		**Bacteroidetes**
Indibacter alkaliphilus LW1		**Bacteroidetes**
Epilithonimonas tenax DSM 16811		**Bacteroidetes**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Galbibacter sp. ck I2 15		**Bacteroidetes**
Gracilimonas tropica DSM 19535		**Bacteroidetes**
Arcticibacter svalbardensis MN127		**Bacteroidetes**
Chrysiogenes arsenatis DSM 11915		**Chrysiogenetes**
Denitrovibrio acetiphilus DSM 12809		**Deferribacteres**
Fusobacterium nucleatum nucleatum ATCC 25586		**Fusobacteria**
Candidatus Gracilibacteria bacterium CG1_02_38_174		**Gracilibacteria**
Candidatus Vecturithrix granuli		**Modulibacteria**
Nitrospinae bacterium RIFCSPLOWO2_12_FULL_47_7		**Nitrospinae**
Nitrospirae bacterium GWC2_57_9		**Nitrospirae**
Nitrospira defluvii		**Nitrospirae**
Brachyspira murdochii DSM 12563		**Spirochaetes**
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Paris)'		**Spirochaetes**
Thermanaerovibrio acidaminovorans DSM 6589		**Synergistetes**
Thermodesulfatator indicus CIR29812 DSM 15286		**Thermodesulfobacteria**
Kosmotoga olearia TBF 19.5.1		**Thermotogae**
Candidatus Wirthbacteria bacterium CG2 30_54_11		**Wirthbacteria**
Heimdallarchaeota LC2		**Asgard group Candidatus Heimdallarchaeota**
Heimdallarchaeota LC3		**Asgard group Candidatus Heimdallarchaeota**
Lokiarchaeum mirabilis		**Asgard group Candidatus Lokiarchaeota**
Odinarchaeota LCB4		**Asgard group Candidatus Odinarchaeota**
Thorarchaeota AB25		**Asgard group Candidatus Thorarchaeota**
Candidatus Aenigmarchaeota CG_4_10_14_3_um_filter_37_21		**DPANN group Candidatus Aenigmarchaeota**
Candidatus Aenigmarchaeota archaeon CG01_land_8_20_14_3_00_37_9		**DPANN group Candidatus Aenigmarchaeota**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Candidatus Diapherotrites archaeon CG11_big_fil_rev_8_21_14_0_20_37_9		**DPANN group Candidatus Diapherotrite**
archaeon GW2011_AR10 (Diapherotrites archaeon AR10?)		**DPANN group Candidatus Diapherotrite**
Candidatus Huberarchaea crystalense		**DPANN group Candidatus Huberarchaea**
Candidatus Micrarchaeum acidiphilum ARMAN-2		**DPANN group Candidatus Micrarchaeota**
Candidatus Micrarchaeota archaeon A_DKE		**DPANN group Candidatus Micrarchaeota**
Candidatus Micrarchaeota archaeon Mia14		**DPANN group Candidatus Micrarchaeota**
Candidatus Nanosalina		**DPANN group Candidatus Nanohaloarchaeota**
Candidatus Nanosalinarum (sp. J07AB56)		**DPANN group Candidatus Nanohaloarchaeota**
Candidatus Nanopetramus SG9		**DPANN group Candidatus Nanohaloarchaeota**
archaeon GW2011_AR13		**DPANN group Candidatus Pacearchaeota**
archaeon GW2011_AR1		**DPANN group Candidatus Pacearchaeota**
archaeon GW2011_AR6		**DPANN group Candidatus Pacearchaeota**
Candidatus Parvarchaeum acidiphilum ARMAN-4		**DPANN group Candidatus Parvarchaeota**
Candidatus Parvarchaeum acidophilus ARMAN-5		**DPANN group Candidatus Parvarchaeota**
archaeon GW2011_AR15		**DPANN group Candidatus Woesearchaeota**
archaeon GW2011_AR20		**DPANN group Candidatus Woesearchaeota**
archaeon GW2011_AR4		**DPANN group Candidatus Woesearchaeota**
Nanoarchaeum equitans		**DPANN group Nanoarchaeota**
Candidatus Nanobsidianus stetteri		**DPANN group Nanoarchaeota**
Candidatus Nanopusillus acidilobi		**DPANN group Nanoarchaeota**
Candidatus Altiarchaeum CG_4_9_14_0_8_um_filter_32_206		**DPANN group Candidatus Altiarchaeales**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
CG_SM1 (Alt 1) - Candidatus Altiarchaeum hamiconexum		**DPANN group Candidatus Altiarchaeales**
WOR_SM1_SCG (Alt 2)		**DPANN group Candidatus Altiarchaeales**
MSI_SMI (Alt 1) - Candidatus Altiarchaeum hamiconexum		**DPANN group Candidatus Altiarchaeales**
Candidatus Altiarchaeales ex4484_2		**DPANN group Candidatus Altiarchaeales**
Aciduliprofundum boonei (T469)		**Euryarchaeota Aciduliprofundum boonei**
Archaeoglobus fulgidus (DSM 4304)		**Euryarchaeota Archaeoglobi**
Archaeoglobus sulfaticallidus		**Euryarchaeota Archaeoglobi**
Hadesarchaea archaeon YNP_N21		**Euryarchaeota Hadesarchaea**
Hadesarchaea YNP_45		**Euryarchaeota Hadesarchaea**
Natronomonas pharaonis (DSM 2160)		**Euryarchaeota Halobacteria**
Haloarcula marismortui		**Euryarchaeota Halobacteria**
Halobacterium salinarum NRC-1		**Euryarchaeota Halobacteria**
Halomicrobium mukohataei		**Euryarchaeota Halobacteria**
Haloferax volcanii		**Euryarchaeota Halobacteria**
Natronococcus amylolyticus		**Euryarchaeota Halobacteria**
Natronobacterium gregoryi (SP2)		**Euryarchaeota Halobacteria**
Methanobacterium bryantii		**Euryarchaeota Methanobacteria**
Methanobacterium formicicum		**Euryarchaeota Methanobacteria**
Methanobacterium sp. 42_16		**Euryarchaeota Methanobacteria**
Methanobacterium subterraneum		**Euryarchaeota Methanobacteria**
Methanothermobacter marburgensis		**Euryarchaeota Methanobacteria**
Methanothermobacter thermautotrophicus		**Euryarchaeota Methanobacteria**
Methanobrevibacter arboriphilus		**Euryarchaeota Methanobacteria**
Methanobrevibacter cuticularis		**Euryarchaeota Methanobacteria**
Methanobrevibacter filiformis		**Euryarchaeota Methanobacteria**
Methanobrevibacter oralis JMR01		**Euryarchaeota Methanobacteria**
Methanobrevibacter millerae		**Euryarchaeota Methanobacteria**
Methanobrevibacter ruminantium		**Euryarchaeota Methanobacteria**
Methanobrevibacter smithii		**Euryarchaeota Methanobacteria**
Methanosphaera cuniculi		**Euryarchaeota Methanobacteria**
Methanosphaera stadtmanae		**Euryarchaeota Methanobacteria**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Methanosphaera sp. WGK6	Archaea	**Euryarchaeota Methanobacteria**
Methanothermus fervidus		**Euryarchaeota Methanobacteria**
Methanocaldococcus jannaschii		**Euryarchaeota Methanococci**
Methanotorris formicicus		**Euryarchaeota Methanococci**
Methanococcus maripaludis		**Euryarchaeota Methanococci**
Methanothermococcus thermolithotrophicus		**Euryarchaeota Methanococci**
Methanocella arvoryzae (Rice Cluster I)		**Euryarchaeota Methanomicrobia**
Methanocella paludicola (SANAE)		**Euryarchaeota Methanomicrobia**
Methanocorpusculum parvum		**Euryarchaeota Methanomicrobia**
Methanocorpusculum labreanum		**Euryarchaeota Methanomicrobia**
Methanoculleus bourgensis		**Euryarchaeota Methanomicrobia**
Methanospirillum hungatei		**Euryarchaeota Methanomicrobia**
Candidatus Methanoperedens nitroreducens (ANME-2d)		**Euryarchaeota Methanomicrobia**
Methanothrix soehngenii		**Euryarchaeota Methanomicrobia**
Methanothrix thermoacetophila (PT)		**Euryarchaeota Methanomicrobia**
Methanococcoides methylutens		**Euryarchaeota Methanomicrobia**
Methanohalophilus mahii		**Euryarchaeota Methanomicrobia**
Methanosarcina acetivorans		**Euryarchaeota Methanomicrobia**
Methanosarcina barkeri str. Fusaro		**Euryarchaeota Methanomicrobia**
Methanosarcina barkeri CM1		**Euryarchaeota Methanomicrobia**
Methanosarcina mazei		**Euryarchaeota Methanomicrobia**
Methanosarcina spelaei		**Euryarchaeota Methanomicrobia**
Methermicoccus shengliensis		**Euryarchaeota Methanomicrobia**
Methanosarcinales archaeon ex4484_138 (GoM-Arcl)		**Euryarchaeota Methanomicrobia**
Methanosarcinales archaeon ex4572_44 (GoM-Arcl)		**Euryarchaeota Methanomicrobia**
Candidatus Syntrophoarchaeum caldarius (ANME-2 - GoM-Arch87-2)		**Euryarchaeota Methanomicrobia**
ANME-2 cluster archaeon HR1		**Euryarchaeota Methanomicrobia**
ANME-1 cluster archaeon ex4572_4		**Euryarchaeota Methanomicrobia**
Arc I group archaeon U1lsi0528_Bin055		**Euryarchaeota Methanomicrobia**
Methanonatronarchaeum thermophilum		**Euryarchaeota Methanonatronarchaeia**
Methanopyrus kandleri AV19		**Euryarchaeota Methanopyri**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Methanopyrus KOL6		**Euryarchaeota Methanopyri**
Theionarchaea archaeon DG-70-1		**Euryarchaeota Theionarchaea**
Theionarchaea archaeon DG-70		**Euryarchaeota Theionarchaea**
Pyrococcus furiosus		**Euryarchaeota Thermococci**
Pyrococcus abyssi		**Euryarchaeota Thermococci**
Pyrococcus horikoshii		**Euryarchaeota Thermococci**
Thermococcus kodakarensis		**Euryarchaeota Thermococci**
Methanomassiliicoccus luminyensis		**Euryarchaeota Thermoplasmata**
Candidatus Methanomassiliicoccus intestinalis		**Euryarchaeota Thermoplasmata**
Candidatus Methanoplasma termitum		**Euryarchaeota Thermoplasmata**
Acidiplasma aeolicum		**Euryarchaeota Thermoplasmata**
Ferroplasma acidarmanus fer1		**Euryarchaeota Thermoplasmata**
Picrophilus torridus		**Euryarchaeota Thermoplasmata**
Thermoplasma volcanium		**Euryarchaeota Thermoplasmata**
Thermoplasmatales archaeon SCGC AB-539-N05		**Euryarchaeota Thermoplasmata**
Thermoplasmatales archaeon B_DKE		**Euryarchaeota Thermoplasmata**
Cuniculiplasma sp. C_DKE		**Euryarchaeota Thermoplasmata**
uncultured marine group II euryarchaeote		**Euryarchaeota unclassified**
Marine group III euryarchaeote CG-Epi2		**Euryarchaeota unclassified**
MSBL1 archaeon SCGC-AAA259E19		**Euryarchaeota unclassified**
MCG-1		**TACK group Candidatus Bathyarchaeota**
MCG-6		**TACK group Candidatus Bathyarchaeota**
MCG-15		**TACK group Candidatus Bathyarchaeota**
Candidatus Bathyarchaeota archaeon BA2		**TACK group Candidatus Bathyarchaeota**
Geothermarchaeota (ex4572_27)		**TACK group Candidatus Geothermarchaeota**
Candidatus Korarchaeum cryptofilum		**TACK group Candidatus Korarchaeota**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Methanomethylicus mesodigestum V2		**TACK group Candidatus Verstraetearchaeota**
Methanosuratus petracarbonis V4		**TACK group Candidatus Verstraetearchaeota**
Acidilobus saccharovorans		**TACK group Crenarchaeota**
Aeropyrum camini		**TACK group Crenarchaeota**
Aeropyrum pernix (K1)		**TACK group Crenarchaeota**
Hyperthermus butylicus		**TACK group Crenarchaeota**
Ignisphaera aggregans (DSM 17230)		**TACK group Crenarchaeota**
Ignicoccus hospitalis		**TACK group Crenarchaeota**
Ignicoccus islandicus (DSM 13165)		**TACK group Crenarchaeota**
Staphylothermus marinus		**TACK group Crenarchaeota**
Fervidicoccus fontis		**TACK group Crenarchaeota**
Sulfolobus solfataricus		**TACK group Crenarchaeota**
Caldivirga maquilingensis		**TACK group Crenarchaeota**
Pyrobaculum aerophilum (str. IM2)		**TACK group Crenarchaeota**
Thermofilum pendens		**TACK group Crenarchaeota**
Candidatus Caldiarchaeum subterraneum		**TACK group Aigarchaeota**
Cenarchaeum symbiosum A		**TACK group Thaumarchaeota**
Nitrosopumilus koreensis		**TACK group Thaumarchaeota**
Nitrosopumilus maritimus SCM1		**TACK group Thaumarchaeota**
Nitrosopumilus limnia		**TACK group Thaumarchaeota**
Candidatus Nitrosomarinus catalina		**TACK group Thaumarchaeota**
Candidatus Nitrososphaera gargensis		**TACK group Thaumarchaeota**
Candidatus Nitrocosmicus oleophilus		**TACK group Thaumarchaeota**
Candidatus Nitrosocaldus islandicus		**TACK group Thaumarchaeota**
Marine group I thaumarchaeote SCGC RSA3		**TACK group Thaumarchaeota**
Thaumarchaeota archaeon SCGC AB-539-E09		**TACK group Thaumarchaeota**
Thaumarchaeota archaeon BS4 (Candidatus Nitrosocaldus cavascurensis)		**TACK group Thaumarchaeota**
Thaumarchaeota Fn1		**TACK group Thaumarchaeota**
Candidatus Marsarchaeota G2 archaeon BE_D		**unclassified**

Table A1: List of species used in the filtration of eukaryotic proteins from contaminants (cont.)

Species name	Domain	Taxonomy
Candidatus Marsarchaeota G1 archaeon OSP_C		**unclassified**

Table A2: List of the 120 proteins associated with the hydrogenosome surface

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001323739.1	TVAG_005910	P-keilini_4runs_TRINITY_DN667_c0_g2_i1.p1	50S-ribosomal-protein-L2,-putative
XP_001323765.1	TVAG_006170	P-keilini_4runs_TRINITY_DN989_c0_g1_i1.p1	60S-ribosomal-protein-L19,-putative
XP_001323773.1	TVAG_006250	P-keilini_4runs_TRINITY_DN915_c0_g1_i1.p1	30S-ribosomal-protein-S8,-putative
XP_001311950.1	TVAG_008680	P-keilini_4runs_TRINITY_DN1020_c0_g1_i1.p1	tubulin-epsilon-chain,-putative
XP_001582527.1	TVAG_013060	P-keilini_4runs_TRINITY_DN4044_c0_g1_i1.p1	60S-ribosomal-protein-L3,-putative
XP_001582635.1	TVAG_014160	P-keilini_4runs_TRINITY_DN645_c0_g1_i1.p2	60S-ribosomal-protein-L12,-putative
XP_001329886.1	TVAG_015800	P-keilini_4runs_TRINITY_DN460_c0_g1_i2.p1	60S-ribosomal-protein-L23,-putative
XP_001317481.1	TVAG_020040	P-keilini_4runs_TRINITY_DN3914_c0_g3_i1.p1	ribosomal-protein-S9,-putative
XP_001314734.1	TVAG_020480	P-keilini_4runs_TRINITY_DN508_c0_g2_i1.p1	40S-ribosomal-protein-S18,-putative
XP_001300897.1	TVAG_033590	P-keilini_4runs_TRINITY_DN670_c0_g1_i1.p1	40S-ribosomal-protein-S6,-putative
XP_001326922.1	TVAG_038050	P-keilini_4runs_TRINITY_DN414_c0_g1_i2.p1	50S-ribosomal-protein-L24p,-putative
XP_001315155.1	TVAG_040820	P-keilini_4runs_TRINITY_DN16934_c0_g1_i1.p2	40S-ribosomal-protein-S17,-putative
XP_001301101.1	TVAG_041350	P-keilini_4runs_TRINITY_DN3310_c1_g1_i1.p1	30S-40S-ribosomal-protein,-putative
XP_001312328.1	TVAG_043060	P-keilini_4runs_TRINITY_DN163620_c0_g1_i1.p1	fructose-bisphosphate-aldolase,-putative
XP_001315627.1	TVAG_043500	P-keilini_4runs_TRINITY_DN242_c0_g2_i1.p1	enolase,-putative
XP_001314691.1	TVAG_044510	P-keilini_4runs_TRINITY_DN842_c0_g2_i4.p1	heat-shock-protein-70-(HSP70)-4,-putative
XP_001330678.1	TVAG_044560	P-keilini_4runs_TRINITY_DN374_c0_g1_i1.p1	50S-ribosomal-protein-L5p,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001323300.1	TVAG_045010	P-keilini_4runs_TRINITY_DN206053_c0_g1_i1.p1	glucokinase,-putative
XP_001308576.1	TVAG_045340	P-keilini_4runs_TRINITY_DN420_c0_g1_i1.p1	polyadenylate-binding-protein,-putative
XP_001298952.1	TVAG_047460	P-keilini_4runs_TRINITY_DN207414_c0_g1_i1.p1	40S-ribosomal-protein-S3a,-putative
XP_001321104.1	TVAG_051160	P-keilini_4runs_TRINITY_DN182683_c0_g1_i1.p1	60S-acidic-ribosomal-protein-P0,-putative
XP_001306738.1	TVAG_054130	P-keilini_4runs_TRINITY_DN92336_c3_g1_i1.p1	60S-ribosomal-protein-L7,-putative
XP_001330619.1	TVAG_054500	P-keilini_4runs_TRINITY_DN198864_c0_g1_i1.p1	50S-ribosomal-protein-L6p,-putative
XP_001308434.1	TVAG_054830	P-keilini_4runs_TRINITY_DN184234_c0_g1_i1.p1	phosphoglucomutase,-putative
XP_001323508.1	TVAG_061890	P-keilini_4runs_TRINITY_DN161037_c0_g1_i1.p1	60S-ribosomal-protein-L18a,-putative
XP_001323512.1	TVAG_061930	P-keilini_4runs_TRINITY_DN218798_c0_g1_i1.p1	glucose-6-phosphate-isomerase,-putative
XP_001320140.1	TVAG_064640	P-keilini_4runs_TRINITY_DN170420_c1_g1_i4.p1	ribosomal-protein-L5,-putative
XP_001318569.1	TVAG_066030	P-keilini_4runs_TRINITY_DN1889_c0_g1_i1.p1	40S-ribosomal-protein-S8,-putative
XP_001328958.1	TVAG_067400	P-keilini_4runs_TRINITY_DN828_c0_g1_i2.p1	elongation-factor-1-alpha,-putative
XP_001584514.1	TVAG_071920	P-keilini_4runs_TRINITY_DN103_c0_g1_i1.p1	40S-ribosomal-protein-S23,-putative
XP_001321243.1	TVAG_073860	P-keilini_4runs_TRINITY_DN2499_c0_g3_i2.p1	phosphoenolpyruvate-protein-phosphotransferase,-putative
XP_001324843.1	TVAG_074480	P-keilini_4runs_TRINITY_DN283_c0_g1_i2.p1	60S-ribosomal-protein-L10a,-putative
XP_001324856.1	TVAG_074610	P-keilini_4runs_TRINITY_DN1754_c0_g1_i2.p1	60S-ribosomal-protein-L10,-putative
XP_001320891.1	TVAG_079260	P-keilini_4runs_TRINITY_DN9_c0_g1_i2.p1	phosphofructokinase,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001579459.1	TVAG_083260	P-keilini_4runs_TRINITY_DN570_c0_g4_i1.p1	60S-ribosomal-protein-L17,-putative
XP_001318159.1	TVAG_087140	P-keilini_4runs_TRINITY_DN173700_c0_g1_i1.p1	arp2/3,-putative
XP_001304619.1	TVAG_092750	P-keilini_4runs_TRINITY_DN690_c0_g1_i1.p1	glucokinase
XP_001303733.1	TVAG_094720	P-keilini_4runs_TRINITY_DN605_c0_g1_i1.p2	60S-ribosomal-protein-L34,-putative
XP_001321933.1	TVAG_098450	P-keilini_4runs_TRINITY_DN218705_c0_g1_i1.p1	40S-ribosomal-protein-S4,-putative
XP_001303193.1	TVAG_099490	P-keilini_4runs_TRINITY_DN184416_c0_g1_i1.p1	glucose-kinase,-putative
XP_001580437.1	TVAG_101690	P-keilini_4runs_TRINITY_DN165121_c0_g1_i1.p1	60S-ribosomal-protein-L24,-putative
XP_001324008.1	TVAG_106800	P-keilini_4runs_TRINITY_DN1152_c0_g1_i1.p1	30S-ribosomal-protein-S3,-putative
XP_001581388.1	TVAG_110140	P-keilini_4runs_TRINITY_DN167_c0_g3_i5.p2	ubiquitin,-putative
XP_001300910.1	TVAG_112230	P-keilini_4runs_TRINITY_DN79301_c0_g2_i1.p1	60S-ribosomal-protein-L13,-putative
XP_001579013.1	TVAG_113710	P-keilini_4runs_TRINITY_DN54853_c0_g1_i1.p2	phosphoglycerate-mutase,-putative
XP_001324949.1	TVAG_117480	P-keilini_4runs_TRINITY_DN330_c0_g1_i3.p1	30S-ribosomal-protein-S11,-putative
XP_001276845.1	TVAG_119330	P-keilini_4runs_TRINITY_DN19_c0_g6_i1.p1	60S-ribosomal-protein-L21,-putative
XP_001276908.1	TVAG_119970	P-keilini_4runs_TRINITY_DN208907_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001276929.1	TVAG_120180	P-keilini_4runs_TRINITY_DN9217_c0_g1_i1.p1	40S-ribosomal-protein-S10,-putative
XP_001277020.1	TVAG_121100	P-keilini_4runs_TRINITY_DN426_c0_g1_i2.p1	60S-ribosomal-protein-L18,-putative
XP_001322974.1	TVAG_121550	P-keilini_4runs_TRINITY_DN1040_c0_g1_i1.p1	60S-ribosomal-protein-L27a,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001324688.1	TVAG_128790	P-keilini_4runs_TRINITY_DN68345_c0_g1_i2.p1	60S-ribosomal-protein-L4,-putative
XP_001324775.1	TVAG_139320	P-keilini_4runs_TRINITY_DN43315_c0_g1_i1.p1	heat-shock-protein,-putative
XP_001320113.1	TVAG_142440	P-keilini_4runs_TRINITY_DN3544_c0_g1_i2.p1	40S-ribosomal-protein-S2,-putative
XP_001579934.1	TVAG_146910	P-keilini_4runs_TRINITY_DN183802_c0_g1_i1.p1	glyceraldehyde-3-phosphate-dehydrogenase,-putative
XP_001309550.1	TVAG_148950	P-keilini_4runs_TRINITY_DN639_c0_g1_i1.p1	50S-ribosomal-protein-L15e,-putative
XP_001309564.1	TVAG_149090	P-keilini_4runs_TRINITY_DN976_c0_g2_i4.p1	actin,-putative
XP_001303635.1	TVAG_152720	P-keilini_4runs_TRINITY_DN330_c0_g1_i3.p1	40S-ribosomal-protein-S14,-putative
XP_001317545.1	TVAG_153560	P-keilini_4runs_TRINITY_DN358_c0_g1_i4.p1	heat-shock-protein,-putative
XP_001313895.1	TVAG_154680	P-keilini_4runs_TRINITY_DN5492_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001310739.1	TVAG_157940	P-keilini_4runs_TRINITY_DN5492_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001580742.1	TVAG_178000	P-keilini_4runs_TRINITY_DN52_c0_g1_i2.p1	60S-ribosomal-protein-L23a,-putative
XP_001583987.1	TVAG_182370	P-keilini_4runs_TRINITY_DN69_c0_g3_i2.p1	chaperonin-containing-t-complex-protein-1,-gamma-subunit,-tcpg,-putative
XP_001584002.1	TVAG_182520	P-keilini_4runs_TRINITY_DN79301_c0_g2_i1.p1	60S-ribosomal-protein-L13,-putative
XP_001580136.1	TVAG_190450	P-keilini_4runs_TRINITY_DN1019_c0_g1_i1.p1	kakapo,-putative
XP_001320814.1	TVAG_191140	P-keilini_4runs_TRINITY_DN528_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001581272.1	TVAG_192620	P-keilini_4runs_TRINITY_DN206419_c0_g1_i1.p1	actin-depolymerizing-factor,-putative
XP_001308024.1	TVAG_204360	P-keilini_4runs_TRINITY_DN168_c0_g2_i1.p1	malate-dehydrogenase,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001308025.1	TVAG_204370	P-keilini_4runs_TRINITY_DN206053_c0_g1_i1.p1	glucokinase,-putative
XP_001303801.1	TVAG_205910	P-keilini_4runs_TRINITY_DN184234_c0_g1_i1.p1	phosphoglucomutase,-putative
XP_001319786.1	TVAG_212020	P-keilini_4runs_TRINITY_DN5548_c1_g1_i1.p1	transketolase,-putative
XP_001325073.1	TVAG_222040	P-keilini_4runs_TRINITY_DN5492_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001310373.1	TVAG_226870	P-keilini_4runs_TRINITY_DN5492_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001303253.1	TVAG_234160	P-keilini_4runs_TRINITY_DN219244_c0_g1_i1.p1	arp2/3-complex-20-kD-subunit,-putative
XP_001581543.1	TVAG_239310	P-keilini_4runs_TRINITY_DN1019_c0_g2_i2.p1	bollus-pemphigoid-antigen,-putative
XP_001320867.1	TVAG_240050	P-keilini_4runs_TRINITY_DN516_c0_g1_i1.p1	40S-ribosomal-protein-sa,-putative
XP_001323701.1	TVAG_248450	P-keilini_4runs_TRINITY_DN1700_c0_g1_i1.p1	peptidyl-tRNA-hydrolase,-putative
XP_001579239.1	TVAG_253650	P-keilini_4runs_TRINITY_DN168_c0_g2_i1.p1	malate-dehydrogenase,-putative
XP_001323043.1	TVAG_258220	P-keilini_4runs_TRINITY_DN7684_c0_g1_i1.p1	glycosyltransferase,-putative
XP_001317828.1	TVAG_263740	P-keilini_4runs_TRINITY_DN242_c0_g1_i2.p1	enolase,-putative
XP_001313821.1	TVAG_265950	P-keilini_4runs_TRINITY_DN56722_c0_g1_i1.p1	60S-ribosomal-protein-L32,-putative
XP_001579758.1	TVAG_268050	P-keilini_4runs_TRINITY_DN220_c0_g1_i1.p1	phosphoglycerate-kinase,-putative
XP_001330352.1	TVAG_272970	P-keilini_4runs_TRINITY_DN362_c0_g1_i1.p1	40S-ribosomal-protein-S24,-putative
XP_001321781.1	TVAG_276310	P-keilini_4runs_TRINITY_DN217279_c1_g1_i1.p1	starch-branching-enzyme-II,-putative
XP_001321791.1	TVAG_276410	P-keilini_4runs_TRINITY_DN3469_c0_g2_i1.p1	translation-elongation-factor,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001304599.1	TVAG_277390	P-keilini_4runs_TRINITY_DN160592_c0_g1_i1.p2	peptidyl-prolyl-cis-trans-isomerase-A,-ppia,-putative
XP_001582074.1	TVAG_282580	P-keilini_4runs_TRINITY_DN52882_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001582118.1	TVAG_283020	P-keilini_4runs_TRINITY_DN587_c0_g1_i1.p1	initiation-factor-5A,-putative
XP_001330357.1	TVAG_292580	P-keilini_4runs_TRINITY_DN16084_c3_g1_i1.p1	immunophilin,-putative
XP_001307577.1	TVAG_293770	P-keilini_4runs_TRINITY_DN160642_c0_g1_i1.p1	phosphofructokinase,-putative
XP_001315498.1	TVAG_299380	P-keilini_4runs_TRINITY_DN330_c0_g1_i3.p1	30S-ribosomal-protein-S11,-putative
XP_001311055.1	TVAG_300000	P-keilini_4runs_TRINITY_DN163620_c0_g1_i1.p1	fructose-bisphosphate-aldolase,-putative
XP_001304306.1	TVAG_319220	P-keilini_4runs_TRINITY_DN362_c0_g2_i1.p1	40S-ribosomal-protein-S24,-putative
XP_001322282.1	TVAG_329460	P-keilini_4runs_TRINITY_DN242_c0_g2_i1.p1	enolase,-putative
XP_001317525.1	TVAG_336940	P-keilini_4runs_TRINITY_DN690_c0_g1_i1.p1	glucokinase,-putative
XP_001319342.1	TVAG_342830	P-keilini_4runs_TRINITY_DN330_c0_g1_i3.p1	40S-ribosomal-protein-S14/30S-ribosomal-protein-S11,-putative
XP_001328722.1	TVAG_348090	P-keilini_4runs_TRINITY_DN207414_c0_g1_i1.p1	40S-ribosomal-protein-S3a,-putative
XP_001328746.1	TVAG_348330	P-keilini_4runs_TRINITY_DN921_c0_g2_i1.p1	glycogen-phosphorylase,-putative
XP_001326306.1	TVAG_351310	P-keilini_4runs_TRINITY_DN351_c0_g1_i1.p1	plastin,-putative
XP_001308720.1	TVAG_354020	P-keilini_4runs_TRINITY_DN160502_c0_g1_i1.p1	centractin,-putative
XP_001307153.1	TVAG_360700	P-keilini_4runs_TRINITY_DN163620_c0_g1_i1.p1	fructose-bisphosphate-aldolase,-putative
XP_001307814.1	TVAG_370550	P-keilini_4runs_TRINITY_DN670_c0_g1_i1.p1	40S-ribosomal-protein-S6,-putative

Table A2: List of the 120 proteins associated with the hydrogenosome surface (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001305702.1	TVAG_373720	P-keilini_4runs_TRINITY_DN2724_c0_g1_i10.p2	pyruvate-kinase,-putative
XP_001306345.1	TVAG_376130	P-keilini_4runs_TRINITY_DN16874_c0_g1_i1.p3	gelsolin,-putative
XP_001304373.1	TVAG_380910	P-keilini_4runs_TRINITY_DN600_c0_g3_i1.p1	DEAD-box-ATP-dependent-RNA-helicase,-putative
XP_001295243.1	TVAG_381030	P-keilini_4runs_TRINITY_DN382_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001302740.1	TVAG_381690	P-keilini_4runs_TRINITY_DN208349_c0_g1_i1.p1	NAD-dependent-epimerase/dehydratase,-putative
XP_001314248.1	TVAG_383940	P-keilini_4runs_TRINITY_DN220_c0_g1_i1.p1	phosphoglycerate-kinase,-putative
XP_001581778.1	TVAG_391760	P-keilini_4runs_TRINITY_DN450_c0_g1_i1.p2	phosphofructokinase,-putative
XP_001327217.1	TVAG_397250	P-keilini_4runs_TRINITY_DN690_c0_g1_i1.p1	glucokinase,-putative
XP_001328502.1	TVAG_423320	P-keilini_4runs_TRINITY_DN79301_c0_g2_i1.p1	60S-ribosomal-protein-L13,-putative
XP_001325139.1	TVAG_430830	P-keilini_4runs_TRINITY_DN9_c0_g1_i2.p1	phosphofructokinase,-putative
XP_001302863.1	TVAG_435000	P-keilini_4runs_TRINITY_DN16084_c3_g1_i1.p1	conserved-hypothetical-protein
XP_001313948.1	TVAG_442070	P-keilini_4runs_TRINITY_DN184416_c0_g1_i1.p1	glucose-kinase,-putative
XP_001579592.1	TVAG_462920	P-keilini_4runs_TRINITY_DN160642_c0_g1_i1.p1	phosphofructokinase,-putative
XP_001325501.1	TVAG_464120	P-keilini_4runs_TRINITY_DN330_c0_g1_i1.p1	30S-ribosomal-protein-S11,-putative
XP_001325506.1	TVAG_464170	P-keilini_4runs_TRINITY_DN242_c0_g2_i1.p1	enolase,-putative
XP_001322192.1	TVAG_482430	P-keilini_4runs_TRINITY_DN426_c0_g1_i2.p1	60S-ribosomal-protein-L18,-putative
XP_001322532.1	TVAG_491670	P-keilini_4runs_TRINITY_DN207881_c0_g1_i1.p1	malic-enzyme,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001584373.1	TVAG_070500	P-keilini_4runs_TRINITY_DN58243_c4_g1_i1.p1	Rab7g-protein,-putative
XP_001584268.1	TVAG_185900	P-keilini_4runs_TRINITY_DN211195_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001584214.1	TVAG_185340	P-keilini_4runs_TRINITY_DN171252_c0_g1_i1.p1	Rab32,-putative
XP_001584210.1	TVAG_185300	P-keilini_4runs_TRINITY_DN25505_c0_g1_i1.p1	Rab11,-putative
XP_001584188.1	TVAG_185080	P-keilini_4runs_TRINITY_DN175067_c0_g1_i1.p1	ran,-putative
XP_001584134.1	TVAG_183850	P-keilini_4runs_TRINITY_DN5301_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001584128.1	TVAG_183790	P-keilini_4runs_TRINITY_DN568_c0_g1_i1.p1	malic-enzyme,-putative
XP_001584099.1	TVAG_183500	P-keilini_4runs_TRINITY_DN541_c0_g1_i1.p1	succinate-thiokinase-¿-subunit
XP_001584080.1	TVAG_183300	P-keilini_4runs_TRINITY_DN194471_c0_g1_i1.p1	aminotransferase-class-V,-putative
XP_001584012.1	TVAG_182620	P-keilini_4runs_TRINITY_DN13155_c3_g1_i1.p1	nitrate,-fromate,-iron-dehydrogenase,-putative
XP_001583906.1	TVAG_076670	P-keilini_4runs_TRINITY_DN1671_c1_g1_i1.p1	small-GTPase-RAB,-putative
XP_001583890.1	TVAG_076510	P-keilini_4runs_TRINITY_DN175_c0_g1_i1.p1	2-amino-3-ketobutyrate-coenzyme-A-ligase,-putative
XP_001583862.1	TVAG_076230	P-keilini_4runs_TRINITY_DN584_c0_g1_i1.p1	nucleotide-binding-protein,-putative
XP_001583785.1	TVAG_075460	P-keilini_4runs_TRINITY_DN206_c3_g1_i1.p1	Rab5b,-putative
XP_001583584.1	TVAG_036230	P-keilini_4runs_TRINITY_DN26_c4_g1_i1.p1	RAB,-putative
XP_001583562.1	TVAG_036010	P-keilini_4runs_TRINITY_DN160973_c0_g1_i1.p1	A-type-flavoprotein
XP_001583303.1	TVAG_377960	P-keilini_4runs_TRINITY_DN26_c4_g1_i1.p1	Rab8,-putative
XP_001583118.1	TVAG_093060	P-keilini_4runs_TRINITY_DN877_c0_g2_i1.p1	Rabx30-protein,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001583042.1	TVAG_456910	P-keilini_4runs_TRINITY_DN941_c2_g5_i1.p1	Rabx32-protein,-putative
XP_001583028.1	TVAG_456770	P-keilini_4runs_TRINITY_DN419_c0_g1_i1.p2	iron-sulfur-cluster-assembly-protein,-putative
XP_001582848.1	TVAG_249220	P-keilini_4runs_TRINITY_DN953_c0_g3_i1.p1	Rab2,-putative
XP_001582728.1	TVAG_237680	P-keilini_4runs_TRINITY_DN208767_c1_g1_i1.p1	ADP,ATP-carrier-protein,-putative
XP_001582674.1	TVAG_237140	P-keilini_4runs_TRINITY_DN842_c0_g4_i1.p1	heat-shock-protein,-putative
XP_001582391.1	TVAG_198430	P-keilini_4runs_TRINITY_DN195283_c0_g1_i1.p1	14-3-3-protein-sigma,-gamma,-zeta,-beta/alpha,-putative
XP_001582383.1	TVAG_198350	P-keilini_4runs_TRINITY_DN6701_c0_g1_i1.p2	hypothetical-protein
XP_001582360.1	TVAG_198110	P-keilini_4runs_TRINITY_DN688_c0_g3_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001582336.1	TVAG_167250	P-keilini_4runs_TRINITY_DN392_c0_g1_i1.p1	chaperonin,-putative
XP_001582335.1	TVAG_167240	P-keilini_4runs_TRINITY_DN16677_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001582155.1	TVAG_283380	P-keilini_4runs_TRINITY_DN2717_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001581795.1	TVAG_391930	P-keilini_4runs_TRINITY_DN184428_c0_g1_i1.p1	RAB,-putative
XP_001581497.1	TVAG_238830	P-keilini_4runs_TRINITY_DN568_c0_g1_i1.p1	malic-enzyme,-putative
XP_001581254.1	TVAG_192440	P-keilini_4runs_TRINITY_DN941_c3_g1_i1.p1	RAC,-putative
XP_001580948.1	TVAG_402160	P-keilini_4runs_TRINITY_DN21606_c0_g1_i1.p1	guanine-nucleotide-exchange-factor,-putative
XP_001580911.1	TVAG_130330	P-keilini_4runs_TRINITY_DN22962_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001580780.1	TVAG_178380	P-keilini_4runs_TRINITY_DN160711_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001580752.1	TVAG_178100	P-keilini_4runs_TRINITY_DN194385_c0_g1_i1.p1	hypothetical-protein

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001580745.1	TVAG_178030	P-keilini_4runs_TRINITY_DN160251_c0_g1_i1.p1	plasma-membrane-calcium-transporting-ATPase,-putative
XP_001580733.1	TVAG_177910	P-keilini_4runs_TRINITY_DN1508_c0_g1_i1.p1	hypothetical-protein
XP_001580703.1	TVAG_433130	P-keilini_4runs_TRINITY_DN285_c1_g1_i1.p1	heat-shock-protein,-putative
XP_001580655.1	TVAG_432650	P-keilini_4runs_TRINITY_DN198613_c0_g1_i1.p1	nitrogen-fixation-protein-nifu,-putative
XP_001580601.1	TVAG_228780	P-keilini_4runs_TRINITY_DN501_c0_g1_i1.p1	alcohol-dehydrogenase,-putative
XP_001580529.1	TVAG_136740	P-keilini_4runs_TRINITY_DN184428_c0_g1_i1.p1	RAB,-putative
XP_001580481.1	TVAG_136260	P-keilini_4runs_TRINITY_DN26_c5_g1_i1.p1	GTP-binding-protein-ypt10,-putative
XP_001580416.1	TVAG_101480	P-keilini_4runs_TRINITY_DN12181_c0_g1_i1.p1	RAB,-putative
XP_001580394.1	TVAG_101260	P-keilini_4runs_TRINITY_DN13976_c0_g1_i1.p1	Rab32,-putative
XP_001580191.1	TVAG_214300	P-keilini_4runs_TRINITY_DN15514_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001580148.1	TVAG_190580	P-keilini_4runs_TRINITY_DN58662_c0_g1_i1.p1	Clan-MG,-familly-M24,-aminopeptidase-P-like-metallopeptidase
XP_001580142.1	TVAG_190510	P-keilini_4runs_TRINITY_DN4576_c0_g1_i1.p1	GTP-binding-protein-rit,-putative
XP_001580044.1	TVAG_247370	P-keilini_4runs_TRINITY_DN184336_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001579755.1	TVAG_268020	P-keilini_4runs_TRINITY_DN194634_c1_g1_i1.p1	aspartate-aminotransferase,-putative
XP_001579748.1	TVAG_267950	P-keilini_4runs_TRINITY_DN207024_c0_g1_i1.p1	lysyl-tRNA-synthetase,-putative
XP_001579740.1	TVAG_267870	P-keilini_4runs_TRINITY_DN269_c0_g1_i2.p1	malic-enzyme,-putative
XP_001579545.1	TVAG_462450	P-keilini_4runs_TRINITY_DN676_c0_g2_i1.p1	RAB,-putative
XP_001579538.1	TVAG_462370	P-keilini_4runs_TRINITY_DN14322_c0_g1_i3.p1	Rabx19-protein,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001579477.1	TVAG_083440	P-keilini_4runs_TRINITY_DN112439_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001579475.1	TVAG_083420	P-keilini_4runs_TRINITY_DN147688_c0_g2_i1.p1	conserved-hypothetical-protein
XP_001579216.1	TVAG_122960	P-keilini_4runs_TRINITY_DN52162_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001579154.1	TVAG_122340	P-keilini_4runs_TRINITY_DN1085_c1_g2_i1.p1	RAB,-putative
XP_001579147.1	TVAG_122270	P-keilini_4runs_TRINITY_DN3020_c0_g2_i1.p1	RAB,-putative
XP_001579030.1	TVAG_113880	P-keilini_4runs_TRINITY_DN13377_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001579029.1	TVAG_113870	P-keilini_4runs_TRINITY_DN182694_c0_g1_i1.p1	Acetyl-CoA-hydrolase,-putative
XP_001578953.1	TVAG_225930	P-keilini_4runs_TRINITY_DN28069_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001276882.1	TVAG_119710	P-keilini_4runs_TRINITY_DN22730_c0_g1_i1.p1	Clan-ME,-family-M16,-insulinase-like-metallopeptidase
XP_001330794.1	TVAG_320200	P-keilini_4runs_TRINITY_DN3462_c0_g2_i1.p1	Rab5b,-putative
XP_001330618.1	TVAG_054490	P-keilini_4runs_TRINITY_DN201059_c0_g1_i1.p1	Tryptophanase,-putative
XP_001330504.1	TVAG_383530	P-keilini_4runs_TRINITY_DN177099_c0_g1_i1.p1	RAB,-putative
XP_001330829.1	TVAG_217400	P-keilini_4runs_TRINITY_DN211716_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001330450.1	TVAG_047890	P-keilini_4runs_TRINITY_DN626_c0_g1_i1.p1	succinate-thiokinase-a-subunit
XP_001314415.1	TVAG_090060	P-keilini_4runs_TRINITY_DN941_c2_g4_i1.p2	GTP-binding-protein-Rab2,-putative
XP_001330332.1	TVAG_272760	P-keilini_4runs_TRINITY_DN218827_c0_g1_i1.p1	glutathione-reductase,-putative
XP_001330320.1	TVAG_217870	P-keilini_4runs_TRINITY_DN19306_c0_g1_i1.p2	nucleotide-binding-protein,-putative
XP_001330242.1	TVAG_325080	P-keilini_4runs_TRINITY_DN183480_c0_g1_i1.p1	conserved-hypothetical-protein

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001330232.1	TVAG_324980	P-keilini_4runs_TRINITY_DN8882_c0_g2_i2.p2	ATP-synthase,-putative
XP_001330176.1	TVAG_395550	P-keilini_4runs_TRINITY_DN182694_c0_g1_i1.p1	Acetyl-CoA-hydrolase,-putative
XP_001329730.1	TVAG_454230	P-keilini_4runs_TRINITY_DN184377_c0_g1_i1.p1	Rab15,-13,-10,-1,-35,-5,-and,-putative
XP_001329309.1	TVAG_297650	P-keilini_4runs_TRINITY_DN601_c0_g1_i1.p1	grpe-protein,-putative
XP_001329276.1	TVAG_297320	P-keilini_4runs_TRINITY_DN26_c4_g1_i1.p1	small-GTPase-rabi,-putative
XP_001329126.1	TVAG_150540	P-keilini_4runs_TRINITY_DN582433_c4_g1_i1.p1	RAB-2,4,14,-putative
XP_001329022.1	TVAG_447580	P-keilini_4runs_TRINITY_DN695333_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001328863.1	TVAG_434770	P-keilini_4runs_TRINITY_DN1107_c0_g1_i1.p1	Rab5b,-putative
XP_001328770.1	TVAG_348580	P-keilini_4runs_TRINITY_DN1522_c0_g1_i1.p1	Rabx26-protein,-putative
XP_001328523.1	TVAG_423530	P-keilini_4runs_TRINITY_DN4238_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001328415.1	TVAG_340860	P-keilini_4runs_TRINITY_DN12251_c0_g1_i1.p2	Rab7g-protein,-putative
XP_001328285.1	TVAG_278280	P-keilini_4runs_TRINITY_DN26_c1_g1_i1.p2	Rabx22-protein,-putative
XP_001328194.1	TVAG_262210	P-keilini_4runs_TRINITY_DN172140_c0_g1_i1.p1	tricarboxylate-transport-protein,-putative
XP_001328129.1	TVAG_165340	P-keilini_4runs_TRINITY_DN626_c0_g1_i1.p1	succinate-thiokinase-a-subunit
XP_001328031.1	TVAG_159810	P-keilini_4runs_TRINITY_DN953_c0_g5_i1.p2	small-GTPase-RAB,-putative
XP_001328023.1	TVAG_159730	P-keilini_4runs_TRINITY_DN206_c0_g2_i2.p1	Rab78,-putative
XP_001327882.1	TVAG_405730	P-keilini_4runs_TRINITY_DN7495_c0_g1_i1.p1	Rabx37-protein,-putative
XP_001327427.1	TVAG_201980	P-keilini_4runs_TRINITY_DN7689_c0_g1_i1.p1	Rab15,-13,-10,-1,-35,-5,-and,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001327371.1	TVAG_392650	P-keilini_4runs_TRINITY_DN781_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001327338.1	TVAG_392320	P-keilini_4runs_TRINITY_DN199807_c0_g1_i1.p1	groes-chaperonin,-putative
XP_001327242.1	TVAG_019190	P-keilini_4runs_TRINITY_DN75143_c0_g2_i1.p1	chaperone-protein-DNAj,-putative
XP_001327227.1	TVAG_397350	P-keilini_4runs_TRINITY_DN160922_c0_g1_i1.p1	Rab9,-putative
XP_001326968.1	TVAG_038530	P-keilini_4runs_TRINITY_DN196128_c0_g1_i1.p1	ornithine-carbamoyltransferase,-putative
XP_001326958.1	TVAG_038420	P-keilini_4runs_TRINITY_DN5113_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001326942.1	TVAG_038250	P-keilini_4runs_TRINITY_DN900_c0_g1_i1.p2	Rab12,-putative
XP_001326936.1	TVAG_038190	P-keilini_4runs_TRINITY_DN8719_c0_g1_i1.p1	small-GTPase-rabi,-putative
XP_001326908.1	TVAG_461020	P-keilini_4runs_TRINITY_DN3412_c0_g1_i1.p1	ABC-transporter,-putative
XP_001326833.1	TVAG_393370	P-keilini_4runs_TRINITY_DN196009_c0_g1_i1.p1	Rab23,-putative
XP_001326755.1	TVAG_388650	P-keilini_4runs_TRINITY_DN30752_c0_g1_i1.p1	serine-palmitoyltransferase-I,-putative
XP_001326629.1	TVAG_255980	P-keilini_4runs_TRINITY_DN9830_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001326421.1	TVAG_373310	P-keilini_4runs_TRINITY_DN85431_c0_g1_i1.p1	RAB,-putative
XP_001326329.1	TVAG_351540	P-keilini_4runs_TRINITY_DN25815_c0_g1_i1.p1	2,4-dienoyl-CoA-reductase-[NADPH],-putative
XP_001326325.1	TVAG_351500	P-keilini_4runs_TRINITY_DN196009_c0_g1_i1.p1	Rab7,-putative
XP_001326307.1	TVAG_351320	P-keilini_4runs_TRINITY_DN17_c0_g2_i1.p1	purine-nucleoside-phosphorylase,-putative
XP_001326088.1	TVAG_044270	P-keilini_4runs_TRINITY_DN494_c1_g1_i1.p1	RAB,-putative
XP_001326029.1	TVAG_468220	P-keilini_4runs_TRINITY_DN13197_c0_g1_i1.p2	conserved-hypothetical-protein

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001325929.1	TVAG_371800	P-keilini_4runs_TRINITY_DN80221_c1_g1_i1.p1	conserved-hypothetical-protein
XP_001325705.1	TVAG_206500	P-keilini_4runs_TRINITY_DN3171_c0_g1_i1.p1	Hydroxylamine-reductase,-putative
XP_001325649.1	TVAG_424580	P-keilini_4runs_TRINITY_DN175264_c0_g1_i1.p1	mevalonate-kinase,-putative
XP_001325272.1	TVAG_212310	P-keilini_4runs_TRINITY_DN12195_c0_g1_i1.p1	Rabx21-protein,-putative
XP_001325171.1	TVAG_404940	P-keilini_4runs_TRINITY_DN2224_c0_g1_i1.p1	RAB,-putative
XP_001324855.1	TVAG_074600	P-keilini_4runs_TRINITY_DN195601_c0_g1_i1.p1	tyrosine-aminotransferase,-putative
XP_001324836.1	TVAG_074410	P-keilini_4runs_TRINITY_DN1405_c0_g2_i1.p1	Rabx41-protein,-putative
XP_001324773.1	TVAG_139300	P-keilini_4runs_TRINITY_DN111_c0_g1_i1.p1	phosphoenolpyruvate-carboxykinase,-putative
XP_001324770.1	TVAG_139270	P-keilini_4runs_TRINITY_DN26_c1_g1_i1.p2	Rab10,-putative
XP_001324695.1	TVAG_128860	P-keilini_4runs_TRINITY_DN211282_c0_g1_i1.p1	RAB-18,-putative
XP_001324545.1	TVAG_161280	P-keilini_4runs_TRINITY_DN494_c0_g3_i2.p1	Rab32,-putative
XP_001324510.1	TVAG_160930	P-keilini_4runs_TRINITY_DN15873_c0_g1_i1.p1	Periplasmic-[Fe]-hydrogenase,-putative
XP_001324271.1	TVAG_038950	P-keilini_4runs_TRINITY_DN941_c0_g4_i1.p1	Rabx3-protein,-putative
XP_001324262.1	TVAG_038850	P-keilini_4runs_TRINITY_DN184336_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001324197.1	TVAG_271570	P-keilini_4runs_TRINITY_DN50689_c0_g1_i1.p1	equilibrative-nucleoside-transporter,-putative
XP_001324090.1	TVAG_362470	P-keilini_4runs_TRINITY_DN7689_c0_g1_i1.p1	Rab17,-putative
XP_001323999.1	TVAG_106710	P-keilini_4runs_TRINITY_DN726_c0_g3_i1.p1	conserved-hypothetical-protein
XP_001323967.1	TVAG_106390	P-keilini_4runs_TRINITY_DN1297_c0_g1_i1.p1	RAB-36-and,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001323815.1	TVAG_081640	P-keilini_4runs_TRINITY_DN46067_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001323810.1	TVAG_081590	P-keilini_4runs_TRINITY_DN34893_c0_g2_i1.p1	Rab5b,-putative
XP_001323776.1	TVAG_006280	P-keilini_4runs_TRINITY_DN9662_c1_g1_i1.p1	RAB,-putative
XP_001323774.1	TVAG_006260	P-keilini_4runs_TRINITY_DN12195_c0_g3_i1.p1	GTP-binding-protein-ypt10,-putative
XP_001323750.1	TVAG_006020	P-keilini_4runs_TRINITY_DN172145_c0_g1_i1.p1	vacuolar-ATP-synthase-subunit-ac39,-putative
XP_001323626.1	TVAG_379850	P-keilini_4runs_TRINITY_DN1175_c2_g1_i1.p1	Rab32,-putative
XP_001323600.1	TVAG_379590	P-keilini_4runs_TRINITY_DN3503_c0_g1_i1.p2	GTPase_rho,-putative
XP_001323596.1	TVAG_379550	P-keilini_4runs_TRINITY_DN195601_c0_g1_i1.p1	tyrosine-aminotransferase,-putative
XP_001323527.1	TVAG_343980	P-keilini_4runs_TRINITY_DN164738_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001323468.1	TVAG_127520	P-keilini_4runs_TRINITY_DN407_c0_g1_i1.p1	N-ethylmaleimide-reductase,-putative
XP_001323391.1	TVAG_498620	P-keilini_4runs_TRINITY_DN403_c0_g1_i1.p1	ankyrin-repeat-cotaining-protein,-putative
XP_001323367.1	TVAG_498380	P-keilini_4runs_TRINITY_DN66760_c1_g1_i1.p1	Rab20,-putative
XP_001323255.1	TVAG_410350	P-keilini_4runs_TRINITY_DN1857_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001323219.1	TVAG_446990	P-keilini_4runs_TRINITY_DN187988_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001323218.1	TVAG_446980	P-keilini_4runs_TRINITY_DN187988_c0_g1_i1.p1	Rab9,-putative
XP_001323182.1	TVAG_446610	P-keilini_4runs_TRINITY_DN2285_c0_g2_i1.p1	small-GTPase-rabd,-putative
XP_001322993.1	TVAG_121740	P-keilini_4runs_TRINITY_DN171252_c0_g1_i1.p1	Rab5,-putative
XP_001322920.1	TVAG_364210	P-keilini_4runs_TRINITY_DN14322_c0_g1_i1.p1	Rabx37-protein,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001322907.1	TVAG_157610	P-keilini_4runs_TRINITY_DN953_c0_g5_i1.p2	rho4,-putative
XP_001322830.1	TVAG_282070	P-keilini_4runs_TRINITY_DN497_c1_g2_i1.p1	Rab8,-putative
XP_001322821.1	TVAG_281980	P-keilini_4runs_TRINITY_DN197200_c0_g1_i1.p1	centrosomal-protein-of-135-kDa,-putative
XP_001322722.1	TVAG_484100	P-keilini_4runs_TRINITY_DN953_c0_g4_i2.p1	Rabx26-protein,-putative
XP_001322593.1	TVAG_109540	P-keilini_4runs_TRINITY_DN41445_c0_g1_i1.p1	serine-hydroxymethyltransferase,-putative
XP_001322449.1	TVAG_118780	P-keilini_4runs_TRINITY_DN108422_c0_g1_i1.p1	calmodulin,-putative
XP_001322271.1	TVAG_329350	P-keilini_4runs_TRINITY_DN7422_c1_g1_i1.p2	Rab5,-putative
XP_001322074.1	TVAG_259320	P-keilini_4runs_TRINITY_DN676_c0_g1_i2.p1	Rabx18-protein,-putative
XP_001322061.1	TVAG_259190	P-keilini_4runs_TRINITY_DN541_c0_g2_i1.p1	succinate-thiokinase-¿-subunit
XP_001321827.1	TVAG_056480	P-keilini_4runs_TRINITY_DN941_c2_g5_i1.p1	ran,-putative
XP_001321627.1	TVAG_420260	P-keilini_4runs_TRINITY_DN183133_c0_g1_i1.p1	ATP-synthase-beta-subunit,-putative
XP_001321621.1	TVAG_420200	P-keilini_4runs_TRINITY_DN608_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001321561.1	TVAG_395100	P-keilini_4runs_TRINITY_DN1052_c0_g1_i1.p1	Rab17,-putative
XP_001321469.1	TVAG_133030	P-keilini_4runs_TRINITY_DN257_c0_g1_i1.p2	NADH-ubiquinone-oxidoreductase-flavoprotein,-putative
XP_001321321.1	TVAG_230580	P-keilini_4runs_TRINITY_DN688_c0_g3_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001321292.1	TVAG_180430	P-keilini_4runs_TRINITY_DN953_c0_g2_i1.p1	RAB-36-and,-putative
XP_001321289.1	TVAG_180400	P-keilini_4runs_TRINITY_DN173495_c0_g1_i1.p1	Phospholipase-C-precursor,-putative
XP_001320926.1	TVAG_079630	P-keilini_4runs_TRINITY_DN124701_c0_g1_i1.p1	conserved-hypothetical-protein

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001320922.1	TVAG_079570	P-keilini_4runs_TRINITY_DN211774_c0_g1_i1.p1	Rabx19-protein,-putative
XP_001320829.1	TVAG_239660	P-keilini_4runs_TRINITY_DN176_c0_g1_i1.p1	cysteine-desulfurylase,-putative
XP_001320436.1	TVAG_308190	P-keilini_4runs_TRINITY_DN174_c0_g1_i1.p1	ral,-putative
XP_001320125.1	TVAG_064490	P-keilini_4runs_TRINITY_DN235_c0_g1_i1.p1	rubrerythrin,-putative
XP_001319879.1	TVAG_430220	P-keilini_4runs_TRINITY_DN3256_c0_g1_i1.p1	RAB,-putative
XP_001319653.1	TVAG_419720	P-keilini_4runs_TRINITY_DN171937_c0_g1_i1.p1	aspartate-aminotransferase,-putative
XP_001319283.1	TVAG_311860	P-keilini_4runs_TRINITY_DN65569_c0_g2_i1.p1	conserved-hypothetical-protein
XP_001319140.1	TVAG_057110	P-keilini_4runs_TRINITY_DN182581_c0_g1_i1.p1	Clan-SC,-family-S33,-methylesterase-like-serine-peptidase
XP_001319122.1	TVAG_056930	P-keilini_4runs_TRINITY_DN14322_c0_g1_i1.p1	RAB-36-and,-putative
XP_001319084.1	TVAG_340390	P-keilini_4runs_TRINITY_DN842_c0_g4_i1.p1	heat-shock-protein-70-(HSP70)-4,-putative
XP_001319083.1	TVAG_340380	P-keilini_4runs_TRINITY_DN91169_c0_g1_i2.p1	conserved-hypothetical-protein
XP_001319074.1	TVAG_340290	P-keilini_4runs_TRINITY_DN269_c0_g1_i2.p1	malic-enzyme,-putative
XP_001318961.1	TVAG_310250	P-keilini_4runs_TRINITY_DN111_c0_g1_i1.p1	phosphoenolpyruvate-carboxykinase,-putative
XP_001318941.1	TVAG_310050	P-keilini_4runs_TRINITY_DN13155_c3_g1_i1.p1	nitrate,-fromate,-iron-dehydrogenase,-putative
XP_001318893.1	TVAG_211200	P-keilini_4runs_TRINITY_DN953_c0_g8_i1.p1	Rabx21-protein,-putative
XP_001318848.1	TVAG_056190	P-keilini_4runs_TRINITY_DN725_c1_g1_i1.p1	Clan-MH,-family-M20,-peptidase-T-like-metallopeptidase
XP_001318715.1	TVAG_257310	P-keilini_4runs_TRINITY_DN941_c2_g3_i1.p1	Rab21,-putative
XP_001318541.1	TVAG_065750	P-keilini_4runs_TRINITY_DN104585_c0_g1_i1.p1	sucrose-transport-protein,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001318394.1	TVAG_098820	P-keilini_4runs_TRINITY_DN195601_c0_g1_i1.p1	tyrosine-aminotransferase,-putative
XP_001318376.1	TVAG_068130	P-keilini_4runs_TRINITY_DN269_c0_g1_i2.p1	malic-enzyme,-putative
XP_001317955.1	TVAG_100550	P-keilini_4runs_TRINITY_DN144216_c0_g1_i1.p1	glycine-cleavage-system-H-protein,-putative
XP_001317694.1	TVAG_286490	P-keilini_4runs_TRINITY_DN3462_c4_g1_i1.p1	RAB,-putative
XP_001317473.1	TVAG_019960	P-keilini_4runs_TRINITY_DN317_c0_g2_i3.p1	conserved-hypothetical-protein
XP_001317292.1	TVAG_191660	P-keilini_4runs_TRINITY_DN184795_c0_g1_i1.p1	groes-chaperonin,-putative
XP_001317232.1	TVAG_416520	P-keilini_4runs_TRINITY_DN1175_c2_g1_i1.p1	RAB-18,-putative
XP_001317174.1	TVAG_040030	P-keilini_4runs_TRINITY_DN1216_c0_g1_i1.p1	Iron-sulfur-flavoprotein
XP_001317169.1	TVAG_039980	P-keilini_4runs_TRINITY_DN108_c0_g2_i1.p3	superoxide-dismutase,-putative
XP_001317041.1	TVAG_226310	P-keilini_4runs_TRINITY_DN474_c2_g1_i1.p1	conserved-hypothetical-protein
XP_001316822.1	TVAG_233350	P-keilini_4runs_TRINITY_DN50982_c0_g1_i1.p1	Clan-ME,-family-M16,-insulinase-like-metallopeptidase
XP_001316713.1	TVAG_453070	P-keilini_4runs_TRINITY_DN258_c0_g1_i1.p1	Rab2,-putative
XP_001316606.1	TVAG_350580	P-keilini_4runs_TRINITY_DN941_c0_g1_i1.p1	Rab21,-putative
XP_001316281.1	TVAG_203620	P-keilini_4runs_TRINITY_DN392_c0_g1_i1.p1	rubisco-subunit-binding-protein-alpha-subunit,-ruba,-putative
XP_001316041.1	TVAG_454570	P-keilini_4runs_TRINITY_DN1322_c0_g1_i1.p1	Embryonic-protein-DC-8,-putative
XP_001315530.1	TVAG_193770	P-keilini_4runs_TRINITY_DN953_c0_g5_i1.p2	RAB,-putative
XP_001315422.1	TVAG_049830	P-keilini_4runs_TRINITY_DN266_c0_g1_i1.p1	disulfide-oxidoreductase,-putative
XP_001315408.1	TVAG_049690	P-keilini_4runs_TRINITY_DN217534_c0_g1_i1.p1	thiamin-pyrophosphokinase,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001315345.1	TVAG_328940	P-keilini_4runs_TRINITY_DN160822_c0_g1_i1.p1	alcohol-dehydrogenase,-putative
XP_001315025.1	TVAG_008100	P-keilini_4runs_TRINITY_DN1297_c0_g1_i1.p1	Rabx38-protein,-putative
XP_001314995.1	TVAG_370000	P-keilini_4runs_TRINITY_DN12195_c0_g1_i1.p1	Rabx30-protein,-putative
XP_001314846.1	TVAG_260830	P-keilini_4runs_TRINITY_DN5872_c0_g2_i1.p1	conserved-hypothetical-protein
XP_001314811.1	TVAG_254890	P-keilini_4runs_TRINITY_DN688_c0_g3_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001314747.1	TVAG_020610	P-keilini_4runs_TRINITY_DN1975_c0_g1_i1.p2	RAB-18,-putative
XP_001314705.1	TVAG_055550	P-keilini_4runs_TRINITY_DN3462_c0_g1_i1.p1	Rab8,-putative
XP_001314029.1	TVAG_412220	P-keilini_4runs_TRINITY_DN269_c0_g1_i2.p1	malic-enzyme,-putative
XP_001313967.1	TVAG_442270	P-keilini_4runs_TRINITY_DN4169_c0_g1_i1.p1	GTP-binding-protein-yptv3,-putative
XP_001313958.1	TVAG_442170	P-keilini_4runs_TRINITY_DN38635_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001313802.1	TVAG_265760	P-keilini_4runs_TRINITY_DN407_c0_g1_i1.p1	N-ethylmaleimide-reductase,-putative
XP_001313682.1	TVAG_096630	P-keilini_4runs_TRINITY_DN217534_c0_g1_i1.p1	thiamin-pyrophosphokinase,-putative
XP_001313584.1	TVAG_208470	P-keilini_4runs_TRINITY_DN160426_c0_g1_i1.p1	threonyl-tRNA-synthetase,-putative
XP_001313551.1	TVAG_181000	P-keilini_4runs_TRINITY_DN206_c0_g5_i1.p1	RAB-19,-41-and,-putative
XP_001313356.1	TVAG_114310	P-keilini_4runs_TRINITY_DN4486_c0_g1_i1.p1	peroxiredoxins,-prx-1,-prx-2,-prx-3,-putative
XP_001313153.1	TVAG_257780	P-keilini_4runs_TRINITY_DN272_c2_g3_i1.p1	biotin-synthase,-putative
XP_001312991.1	TVAG_196220	P-keilini_4runs_TRINITY_DN207877_c0_g1_i1.p1	protein-brittle-1,-chloroplast-precursor,-putative
XP_001312927.1	TVAG_029020	P-keilini_4runs_TRINITY_DN1080_c0_g1_i1.p1	small-GTPase-rabh,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001312785.1	TVAG_424920	P-keilini_4runs_TRINITY_DN66760_c1_g1_i1.p1	RAB-3-and,-putative
XP_001312753.1	TVAG_088050	P-keilini_4runs_TRINITY_DN392_c0_g1_i1.p1	chaperonin,-putative
XP_001312620.1	TVAG_468600	P-keilini_4runs_TRINITY_DN182200_c0_g1_i1.p1	WD-repeat-protein,-putative
XP_001312469.1	TVAG_241150	P-keilini_4runs_TRINITY_DN26_c1_g1_i1.p2	Rabx24-protein,-putative
XP_001312254.1	TVAG_152430	P-keilini_4runs_TRINITY_DN159673_c0_g1_i1.p1	lysyl-tRNA-synthetase,-putative
XP_001312198.1	TVAG_112840	P-keilini_4runs_TRINITY_DN187988_c0_g1_i1.p1	Rabx15-protein,-putative
XP_001312168.1	TVAG_296220	P-keilini_4runs_TRINITY_DN160340_c0_g1_i1.p1	NADH-dehydrogenase-24-kDa-subunit,-putative
XP_001312142.1	TVAG_236570	P-keilini_4runs_TRINITY_DN4000_c0_g1_i1.p1	Rab32,-putative
XP_001312073.1	TVAG_115470	P-keilini_4runs_TRINITY_DN195996_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001311960.1	TVAG_008790	P-keilini_4runs_TRINITY_DN71584_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001311873.1	TVAG_321030	P-keilini_4runs_TRINITY_DN195918_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001311871.1	TVAG_321010	P-keilini_4runs_TRINITY_DN49090_c0_g2_i1.p1	AMP-dependent-ligase/synthetase,-putative
XP_001311861.1	TVAG_242960	P-keilini_4runs_TRINITY_DN688_c0_g3_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001311772.1	TVAG_409800	P-keilini_4runs_TRINITY_DN207636_c0_g1_i1.p1	RAB,-putative
XP_001311734.1	TVAG_479760	P-keilini_4runs_TRINITY_DN47132_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001311148.1	TVAG_263350	P-keilini_4runs_TRINITY_DN3153_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001311109.1	TVAG_047800	P-keilini_4runs_TRINITY_DN2125_c0_g1_i1.p1	Rab11,-putative
XP_001310420.1	TVAG_335500	P-keilini_4runs_TRINITY_DN17664_c0_g1_i1.p1	conserved-hypothetical-protein

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001310180.1	TVAG_037570	P-keilini_4runs_TRINITY_DN13155_c0_g4_i1.p1	NADH-ubiquinone-oxidoreductase,-putative
XP_001310176.1	TVAG_037530	P-keilini_4runs_TRINITY_DN8689_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001309934.1	TVAG_440200	P-keilini_4runs_TRINITY_DN164467_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001309798.1	TVAG_390750	P-keilini_4runs_TRINITY_DN26408_c0_g3_i1.p2	RAB-18,-putative
XP_001309776.1	TVAG_470110	P-keilini_4runs_TRINITY_DN197003_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001309717.1	TVAG_216900	P-keilini_4runs_TRINITY_DN953_c0_g9_i1.p1	ras,-putative
XP_001309656.1	TVAG_075320	P-keilini_4runs_TRINITY_DN3946_c0_g1_i1.p2	vacuolar-proton-ATPase,-putative
XP_001309521.1	TVAG_018050	P-keilini_4runs_TRINITY_DN217915_c0_g1_i1.p2	RAB-18,-putative
XP_001309408.1	TVAG_277590	P-keilini_4runs_TRINITY_DN160251_c0_g1_i1.p1	cation-transporting-ATPase,-putative
XP_001309398.1	TVAG_024790	P-keilini_4runs_TRINITY_DN494_c0_g1_i1.p2	septum-promoting-GTP-binding-protein,-putative
XP_001309295.1	TVAG_337970	P-keilini_4runs_TRINITY_DN106_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001309218.1	TVAG_264120	P-keilini_4runs_TRINITY_DN34504_c0_g3_i1.p1	pecanex,-putative
XP_001309182.1	TVAG_205390	P-keilini_4runs_TRINITY_DN350_c0_g1_i2.p1	GTPase-mss1/trme,-putative
XP_001308879.1	TVAG_077910	P-keilini_4runs_TRINITY_DN4238_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001308636.1	TVAG_256470	P-keilini_4runs_TRINITY_DN217534_c0_g1_i1.p1	thiamin-pyrophosphokinase,-putative
XP_001308553.1	TVAG_047210	P-keilini_4runs_TRINITY_DN2207_c0_g1_i1.p1	2-hydroxyacid-dehydrogenase,-putative
XP_001308528.1	TVAG_459470	P-keilini_4runs_TRINITY_DN1975_c0_g1_i1.p2	Rab9,-putative
XP_001308237.1	TVAG_032090	P-keilini_4runs_TRINITY_DN622_c0_g1_i1.p1	Co-chaperone-protein-HscB-Hsc20,-mitochondrial-precursor,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001308201.1	TVAG_051830	P-keilini_4runs_TRINITY_DN3462_c4_g1_i1.p1	RAC,-putative
XP_001308200.1	TVAG_051820	P-keilini_4runs_TRINITY_DN172140_c0_g1_i1.p1	tricarboxylate-transport-protein,-putative
XP_001308167.1	TVAG_336320	P-keilini_4runs_TRINITY_DN3171_c0_g2_i1.p1	Hydroxylamine-reductase,-putative
XP_001307912.1	TVAG_048600	P-keilini_4runs_TRINITY_DN1721_c0_g1_i1.p1	small-GTPase-rabi,-putative
XP_001307775.1	TVAG_049140	P-keilini_4runs_TRINITY_DN108_c0_g2_i1.p3	superoxide-dismutase-[fe],-putative
XP_001307690.1	TVAG_346230	P-keilini_4runs_TRINITY_DN41178_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001307488.1	TVAG_300910	P-keilini_4runs_TRINITY_DN199171_c0_g1_i1.p1	Rabx18-protein,-putative
XP_001307405.1	TVAG_458060	P-keilini_4runs_TRINITY_DN157931_c0_g2_i1.p1	conserved-hypothetical-protein
XP_001307320.1	TVAG_194760	P-keilini_4runs_TRINITY_DN64940_c0_g1_i1.p1	guanine-nucleotide-exchange-factor,-putative
XP_001307251.1	TVAG_466790	P-keilini_4runs_TRINITY_DN206198_c0_g1_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001307088.1	TVAG_105770	P-keilini_4runs_TRINITY_DN206198_c0_g1_i1.p1	pyruvate-flavodoxin-oxidoreductase,-putative
XP_001307064.1	TVAG_080400	P-keilini_4runs_TRINITY_DN953_c0_g10_i1.p1	Rabx31-protein,-putative
XP_001306984.1	TVAG_151010	P-keilini_4runs_TRINITY_DN2353_c0_g1_i3.p3	Rabx19-protein,-putative
XP_001306776.1	TVAG_224980	P-keilini_4runs_TRINITY_DN160002_c0_g1_i1.p1	Clan-MH,-family-M20,-peptidase-T-like-metallopeptidase
XP_001306669.1	TVAG_455090	P-keilini_4runs_TRINITY_DN4238_c0_g1_i1.p2	conserved-hypothetical-protein
XP_001306447.1	TVAG_132350	P-keilini_4runs_TRINITY_DN830_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001306356.1	TVAG_104710	P-keilini_4runs_TRINITY_DN1085_c0_g1_i1.p1	Rabx3-protein,-putative
XP_001306230.1	TVAG_277050	P-keilini_4runs_TRINITY_DN27394_c0_g1_i1.p1	Citrate-lyase-beta-chain,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001305871.1	TVAG_489800	P-keilini_4runs_TRINITY_DN222480_c0_g1_i1.p1	NADH-dehydrogenase-51-kDa-subunit,-putative
XP_001305709.1	TVAG_361590	P-keilini_4runs_TRINITY_DN13155_c0_g4_i1.p1	NADH-ubiquinone-oxidoreductase,-putative
XP_001305704.1	TVAG_361540	P-keilini_4runs_TRINITY_DN419_c0_g1_i1.p2	iron-sulfur-assembly-protein,-putative
XP_001305570.1	TVAG_349870	P-keilini_4runs_TRINITY_DN26_c0_g1_i2.p1	Rabx15-protein,-putative
XP_001305426.1	TVAG_344520	P-keilini_4runs_TRINITY_DN914_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001305403.1	TVAG_331490	P-keilini_4runs_TRINITY_DN9662_c1_g1_i1.p1	Rabx18-protein,-putative
XP_001305368.1	TVAG_327470	P-keilini_4runs_TRINITY_DN771_c0_g1_i1.p1	alcohol-dehydrogenase,-putative
XP_001305213.1	TVAG_262750	P-keilini_4runs_TRINITY_DN160677_c0_g1_i1.p1	vacuolar-ATP-synthase-subunit-H,-putative
XP_001305174.1	TVAG_164890	P-keilini_4runs_TRINITY_DN182694_c0_g1_i1.p1	Acetyl-CoA-hydrolase,-putative
XP_001305114.1	TVAG_386080	P-keilini_4runs_TRINITY_DN170837_c0_g1_i1.p1	Clan-MG,-familly-M24,-aminopeptidase-P-like-metallopeptidase
XP_001305092.1	TVAG_383350	P-keilini_4runs_TRINITY_DN196009_c0_g1_i1.p1	RAB-2,4,14,-putative
XP_001304655.1	TVAG_384490	P-keilini_4runs_TRINITY_DN429_c0_g1_i1.p1	Rab9,-putative
XP_001304618.1	TVAG_092740	P-keilini_4runs_TRINITY_DN219560_c0_g1_i1.p1	small-GTPase-rabi,-putative
XP_001304246.1	TVAG_065320	P-keilini_4runs_TRINITY_DN2353_c0_g1_i3.p3	Rab21,-putative
XP_001304133.1	TVAG_220970	P-keilini_4runs_TRINITY_DN16958_c0_g1_i1.p2	RAB,-putative
XP_001304067.1	TVAG_147840	P-keilini_4runs_TRINITY_DN210433_c0_g1_i1.p1	Rab21,-putative
XP_001304062.1	TVAG_147790	P-keilini_4runs_TRINITY_DN6516_c0_g2_i1.p1	cysteine/methionine-metabolism-pyridoxal-5-phosphate-enzymes,-putative
XP_001303981.1	TVAG_144730	P-keilini_4runs_TRINITY_DN541_c0_g1_i1.p1	succinate-thiokinase-¿-subunit

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001303783.1	TVAG_445730	P-keilini_4runs_TRINITY_DN199807_c0_g1_i1.p1	groes-chaperonin,-putative
XP_001303674.1	TVAG_092170	P-keilini_4runs_TRINITY_DN1127_c0_g1_i1.p1	preprotein-translocase-secy-subunit,-putative
XP_001303658.1	TVAG_134510	P-keilini_4runs_TRINITY_DN219911_c0_g1_i1.p1	RAB-18,-putative
XP_001303641.1	TVAG_328110	P-keilini_4runs_TRINITY_DN26_c0_g1_i2.p1	RAB-18,-putative
XP_001303632.1	TVAG_152690	P-keilini_4runs_TRINITY_DN407_c0_g1_i1.p1	N-ethylmaleimide-reductase,-putative
XP_001303268.1	TVAG_385350	P-keilini_4runs_TRINITY_DN183936_c0_g1_i1.p1	thioredoxin,-putative
XP_001303252.1	TVAG_234150	P-keilini_4runs_TRINITY_DN109943_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001303146.1	TVAG_415960	P-keilini_4runs_TRINITY_DN3256_c0_g1_i1.p1	Rab9,-putative
XP_001303059.1	TVAG_088220	P-keilini_4runs_TRINITY_DN195601_c0_g1_i1.p1	aspartate-aminotransferase,-putative
XP_001302997.1	TVAG_371280	P-keilini_4runs_TRINITY_DN12181_c0_g1_i1.p1	Rab2,-putative
XP_001302917.1	TVAG_124590	P-keilini_4runs_TRINITY_DN4394_c0_g1_i1.p1	Rab6,-putative
XP_001302913.1	TVAG_124540	P-keilini_4runs_TRINITY_DN158973_c0_g1_i1.p1	GTPase_rho,-putative
XP_001302832.1	TVAG_320300	P-keilini_4runs_TRINITY_DN184375_c0_g1_i1.p1	GTP-binding-protein-Rab2,-putative
XP_001302560.1	TVAG_386000	P-keilini_4runs_TRINITY_DN159371_c0_g1_i1.p1	Receptor-expression-enhancing-protein,-putative
XP_001302496.1	TVAG_356810	P-keilini_4runs_TRINITY_DN144134_c0_g2_i1.p1	NimA-like-protein
XP_001302364.1	TVAG_327760	P-keilini_4runs_TRINITY_DN1216_c0_g1_i1.p1	Iron-sulfur-flavoprotein
XP_001302013.1	TVAG_261280	P-keilini_4runs_TRINITY_DN2117_c0_g1_i1.p1	GTP-binding-protein-ypt11,-putative
XP_001301760.1	TVAG_025980	P-keilini_4runs_TRINITY_DN11926_c1_g1_i2.p1	glutamate-dehydrogenase,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001301244.1	TVAG_158270	P-keilini_4runs_TRINITY_DN222101_c0_g1_i1.p1	RAB,-putative
XP_001301100.1	TVAG_041340	P-keilini_4runs_TRINITY_DN199807_c0_g1_i1.p1	groes-chaperonin,-putative
XP_001301044.1	TVAG_422690	P-keilini_4runs_TRINITY_DN1052_c0_g1_i1.p1	Rab9,-putative
XP_001301038.1	TVAG_422630	P-keilini_4runs_TRINITY_DN622_c0_g1_i1.p1	Co-chaperone-protein-HscB-Hsc20,-mitochondrial-precursor,-putative
XP_001300804.1	TVAG_293370	P-keilini_4runs_TRINITY_DN207058_c0_g1_i1.p1	nucleoside-diphosphate-kinase,-putative
XP_001300601.1	TVAG_001130	P-keilini_4runs_TRINITY_DN53254_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001300482.1	TVAG_318670	P-keilini_4runs_TRINITY_DN626_c0_g1_i1.p1	succinate-thiokinase-a-subunit
XP_001300294.1	TVAG_015270	P-keilini_4runs_TRINITY_DN6278_c0_g1_i2.p2	small-GTPase-rabh,-putative
XP_001300248.1	TVAG_169740	P-keilini_4runs_TRINITY_DN219911_c0_g1_i1.p1	RAB-19,-41-and,-putative
XP_001299687.1	TVAG_440690	P-keilini_4runs_TRINITY_DN26_c4_g1_i1.p1	small-GTPase-rabh,-putative
XP_001299604.1	TVAG_527180	P-keilini_4runs_TRINITY_DN327_c2_g1_i1.p1	RAB,-putative
XP_001299513.1	TVAG_177600	P-keilini_4runs_TRINITY_DN183618_c0_g1_i1.p1	glycine-cleavage-system-H-protein,-putative
XP_001299483.1	TVAG_450220	P-keilini_4runs_TRINITY_DN96108_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001299482.1	TVAG_399860	P-keilini_4runs_TRINITY_DN71456_c0_g1_i1.p1	Ferredoxin-2
XP_001299372.1	TVAG_499340	P-keilini_4runs_TRINITY_DN497_c1_g1_i1.p1	hypothetical-protein
XP_001299219.1	TVAG_528800	P-keilini_4runs_TRINITY_DN3462_c4_g1_i1.p1	RAB-2,4,14,-putative
XP_001299204.1	TVAG_450060	P-keilini_4runs_TRINITY_DN13377_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001298987.1	TVAG_126970	P-keilini_4runs_TRINITY_DN207052_c0_g1_i1.p1	RAB-GDP-dissociation-inhibitor,-putative

Table A3: List of 333 putative membrane proteins, enzymes and hypothetical proteins identified in the hydrogenosome (cont.)

T. vaginalis **protein accession number**	***T. vaginalis*** **gene_ID**	***P. keilini*** **protein header**	**Protein function**
XP_001298262.1	TVAG_022530	P-keilini_4runs_TRINITY_DN30686_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001297863.1	TVAG_377380	P-keilini_4runs_TRINITY_DN1554_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001297704.1	TVAG_085320	P-keilini_4runs_TRINITY_DN2659_c0_g1_i1.p1	small-GTPase-rabi,-putative
XP_001297366.1	TVAG_530140	P-keilini_4runs_TRINITY_DN160991_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001297292.1	TVAG_504530	P-keilini_4runs_TRINITY_DN211774_c0_g1_i1.p1	Rabx26-protein,-putative
XP_001296818.1	TVAG_060820	P-keilini_4runs_TRINITY_DN1085_c1_g1_i1.p1	RAB,-putative
XP_001296212.1	TVAG_082020	P-keilini_4runs_TRINITY_DN159740_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001294582.1	TVAG_547520	P-keilini_4runs_TRINITY_DN218014_c0_g1_i1.p1	threonine-synthase,-putative
XP_001294517.1	TVAG_416100	P-keilini_4runs_TRINITY_DN568_c0_g1_i1.p1	malic-enzyme,-putative

Table A4: List of 32 proteins that are probable contaminants to the hydrogenosome

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001309030.1	TVAG_014920	P-keilini_4runs_TRINITY_DN816_c0_g1_i1.p1	histone-H4,-putative
XP_001581184.1	TVAG_021440	P-keilini_4runs_TRINITY_DN767_c0_g1_i11.p1	histone-H2a,-putative
XP_001326531.1	TVAG_026290	P-keilini_4runs_TRINITY_DN5367_c0_g1_i1.p1	oxysterol-binding-protein,-putative
XP_001326541.1	TVAG_026390	P-keilini_4runs_TRINITY_DN895_c1_g1_i1.p1	histone-H2b,-putative
XP_001324277.1	TVAG_039020	P-keilini_4runs_TRINITY_DN185383_c0_g1_i1.p1	conserved-hypothetical-protein
XP_001315624.1	TVAG_043470	P-keilini_4runs_TRINITY_DN185187_c0_g1_i1.p1	RNA-binding-protein,-putative
XP_001326085.1	TVAG_044240	P-keilini_4runs_TRINITY_DN1150_c2_g1_i1.p1	adaptin,-alpha/gamma/epsilon,-putative
XP_001310504.1	TVAG_064150	P-keilini_4runs_TRINITY_DN326_c0_g1_i1.p1	ADP-ribosylation-factor,-arf,-putative
XP_001584436.1	TVAG_071140	P-keilini_4runs_TRINITY_DN1695_c0_g1_i1.p1	clathrin-coat-assembly-protein-ap17,-putative
XP_001302899.1	TVAG_091590	P-keilini_4runs_TRINITY_DN220031_c0_g1_i1.p1	signal-recognition-particle-68-kD-protein,-putative
XP_001312123.1	TVAG_092490	P-keilini_4runs_TRINITY_DN2911_c0_g3_i1.p1	heat-shock-protein-70kD,-putative
XP_001305631.1	TVAG_148040	P-keilini_4runs_TRINITY_DN326_c0_g1_i1.p1	ADP-ribosylation-factor,-arf,-putative
XP_001329089.1	TVAG_150170	P-keilini_4runs_TRINITY_DN172307_c0_g1_i1.p1	ormdl-proteins,-putative
XP_001323895.1	TVAG_158990	P-keilini_4runs_TRINITY_DN185187_c0_g1_i1.p1	Heterogeneous-nuclear-ribonucleoprotein-A/B,-putative
XP_001315127.1	TVAG_169060	P-keilini_4runs_TRINITY_DN182708_c0_g1_i1.p1	GTPase_rho,-putative
XP_001322736.1	TVAG_184150	P-keilini_4runs_TRINITY_DN167_c0_g3_i4.p1	ubiquitin,-putative

Table A4: List of 32 proteins that are probable contaminants to the hydrogenosome (cont.)

T. vaginalis protein accession number	*T. vaginalis* gene_ID	*P. keilini* protein header	Protein function
XP_001319610.1	TVAG_204940	P-keilini_4runs_TRINITY_DN144_c0_g1_i1.p1	calreticulin-and-calnexin,-putative
XP_001310131.1	TVAG_218150	P-keilini_4runs_TRINITY_DN1148_c0_g2_i2.p1	adenylate-cyclase,-putative
XP_001582834.1	TVAG_249080	P-keilini_4runs_TRINITY_DN160305_c0_g1_i1.p1	snare-proteins,-putative
XP_001317755.1	TVAG_318870	P-keilini_4runs_TRINITY_DN364_c0_g1_i1.p1	spermatogenesis-associated-factor,-putative
XP_001311074.1	TVAG_332540	P-keilini_4runs_TRINITY_DN121152_c0_g1_i4.p1	DNA-repair-protein-rad50,-putative
XP_001315753.1	TVAG_369020	P-keilini_4runs_TRINITY_DN1155_c0_g1_i1.p1	clathrin-heavy-chain,-putative
XP_001315754.1	TVAG_369030	P-keilini_4runs_TRINITY_DN1155_c0_g1_i1.p1	clathrin-heavy-chain,-putative
XP_001325917.1	TVAG_371680	P-keilini_4runs_TRINITY_DN164153_c0_g1_i1.p1	cop9-signalosome-complex-subunit,-putative
XP_001583432.1	TVAG_379250	P-keilini_4runs_TRINITY_DN83024_c0_g2_i1.p1	UDP-glucose-glycoprotein: glucosyltransferase,-putative
XP_001323610.1	TVAG_379690	P-keilini_4runs_TRINITY_DN1276_c1_g1_i1.p1	clathrin-coat-assembly-protein,-putative
XP_001299484.1	TVAG_450230	P-keilini_4runs_TRINITY_DN207646_c0_g1_i1.p1	clathrin-coat-associated-protein-ap-50,-putative
XP_001325490.1	TVAG_464010	P-keilini_4runs_TRINITY_DN144_c0_g1_i1.p1	calreticulin,-putative
XP_001322527.1	TVAG_491610	P-keilini_4runs_TRINITY_DN13973_c1_g1_i1.p1	serine-threonine-protein-kinase,-putative
XP_001295063.1	TVAG_532880	P-keilini_4runs_TRINITY_DN1155_c0_g1_i1.p1	clathrin-heavy-chain,-putative
XP_001295192.1	TVAG_547230	P-keilini_4runs_TRINITY_DN182595_c0_g1_i1.p1	adaptin,-alpha/gamma/epsilon,-putative
XP_001294813.1	TVAG_562550	P-keilini_4runs_TRINITY_DN1155_c0_g1_i1.p1	clathrin-heavy-chain,-putative